Estimation of a logistic regression model by a genetic algorithm to predict pipe failures in sewer networks

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

Sewer networks are mainly composed of pipelines which are in charge of transporting sewage and rainwater to wastewater treatment plants. Breakages of sewer pipes may have many negative consequences, such as accidents, flooding, pollution or extra costs. Machine learning arises as a very powerful tool to predict these incidents when the amount of available data is large enough. In this study, a real-coded genetic algorithm is implemented to estimate the optimal weights of a logistic regression model whose objective is to forecast pipe failures within a wastewater network. On the one hand, logistic regression has a great applicability to classification problems, specifically for those whose output variable is binary. On the other hand, the genetic algorithm is a bio-inspired technique which explores the search space to find the optimal or near-optimal solution.

Firstly, the historical database is transformed to a yearly basis and the model is estimated with a set of years (training data). Thus, the classifier can assign a probability of failure to each network pipe, which will allow distinguishing those pipes more prone to fail. Finally, the performance of the model is assessed with unseen data (test data).

To check its performance, the methodology is applied to a real database of a Spanish city. Results demonstrate that if 3% of pipe segments had been replaced, whose failure probabilities were higher than 0.75, more than 25% of unexpected pipe failures could have been prevented.

Keywords: Logistic regression; Binary classifier; Pipe failures; Genetic Algorithm; Sewer networks

1. Introduction

Once water is used by humans or industries, it needs to be adequately drawn and treated in order to close the water cycle. A sewer network ensures the collection of wastewater and transports it to wastewater treatment plants.

In general, sewer networks require two types of actions according to their nature: preventive and corrective actions. Preventive actions are mainly cleanings while corrective actions are unblocking or pipe replacements due to defects or breakages. The cost of corrective actions can be two to ten times higher than preventive actions (Anbari, Tabesh, & Roozbahani, 2017). Consequently, if these unexpected incidences, which induce corrective actions, were

forecasted, significant costs would be saved. Moreover, an unexpected pipe breakage in a sewer network can cause environmental damage, such as pollution or flooding.

Monitored visual inspections have commonly been carried out to detect incidents inside sewer pipes. Nowadays, sophisticated techniques are emerging to automatically identify these defects through image processing. Artificial intelligence methods, as deep convolutional neural networks (Hassan et al., 2019; Li, Cong, & Guo, 2019) or support vector machines (Halfawy & Hengmeechai, 2014; M. Der Yang & Su, 2009; M. D. Yang & Su, 2008), are used for image classification in order to determine defects or incidences. However, visual inspections are time-consuming and networks usually have an extensive total length. Therefore, it is only possible to make daily tracking of certain network pipes. For this reason, it is important to know in advance which parts of the network are more vulnerable in order to prioritise the inspections of these areas. Logistic regression arises as a method capable of determining the probability of suffering a defect of each pipe. Therefore, it might be a starting point for an efficient inspections' planning. One previous work (Sousa, Matos, & Matias, 2014) already used this model for the same purpose, obtaining fairly accurate predictions for a small size network. Furthermore, studies whose objective is to predict pipe failures of water supply networks can also be found, as (Kleiner & Rajani, 2012; Yamijala, Guikema, & Brumbelow, 2009).

Other applied methodology is evolutionary polynomial regression (EPR) whose solutions are mathematical equations that represent variables' relations (Kleiner & Rajani, 2001; Savic et al., 2006; Ugarelli, Kristensen, Røstum, Sægrov, & Di Federico, 2009). This method only has sense when the number of explanatory variables is limited. Its purpose is usually descriptive rather than predictive, focusing on the physical properties' influence on pipe failures. Both neural networks and support vector machines are well-known to make precise predictions, it has been demonstrated in various researches (Khan, Zayed, & Moselhi, 2009; Mashford, Marlow, Tran, & May, 2011). The major disadvantage of these techniques is that they do not allow users to analyse the role that each variable has in the predictions. In fact, they are usually referred to as black box techniques.

For training a predictive model, historical data is necessary and it must be reliable. Most failure predictive researches in the area have been focused on water supply network while there are fewer studies focused on analysing pipe failures of sewer networks. This entails the information about which factors are the most influential in sewer pipe failures to be scarce. In (Anbari et al., 2017), incomplete data is combined with expert opinions for risk assessment of sewer networks using Bayesian Networks. In (Kuliczkowska, 2016), eleven factors are identified which affect failures caused by internal corrosion in concrete pipes: pipe diameter, depth of the pipe, type of soil, sewer function, road type, traffic and several consequence factors as the environmental impact. Actually, in (Younis & Knight, 2010), it was found that the deterioration of reinforced concrete pipes is age-related due to the corrosion. On the contrary, it does not happen in vitrified clay pipes.

It is commonly accepted that most incidences are caused by blockages (Bailey et al., 2015). According to (Savic et al., 2006), these have a direct correlation with pipe diameter and an inverse correlation with the pipe's length. The factors pipe age and type of water were also identified by (Ugarelli et al., 2009) as the most important factors for the appearance of blockages.

This paper provides a useful and self-explanatory methodology to determine the risk of failure of sewer pipes, such as blockages and breakages. The parameters of a logistic regression model are estimated using the well-known genetic algorithm. Moreover, the proposed methodology is tested with real network data. In section 2, a detailed explanation of the two methods employed is presented. Section 3 describes the case study and, specifically, those explanatory variables which have been included in the model. Section 4 shows and discusses the obtained results. Finally, conclusions are presented in section 5.

2. Proposed methodology

Logistic regression (LR) is the model used in this study to predict pipe failures in sewer networks. Its training consists of estimating certain weights to maximise the log-likelihood function. To achieve such goal, a non-linear model must be solved. There are several methods capable to solve this type of model, such as the Newton-Rapson method or the gradient-descent algorithm. Both of them update the weights iteratively, using derivatives, until a minimum or maximum is found. The major disadvantage is that they are computationally expensive when the objective function is complex, and, in addition, for those cases with multiple extrema, they may not converge to the optimum. Other approach could be to estimate the weights using the binomial boosting algorithm which seems to be more appropriate when there are several covariables and noisy data (de Menezes, Liska, Cirillo, & Vivanco, 2017).

Genetic algorithms (GAs) arise as bio-inspired techniques which enable to explore the search space, increasing the chances of finding the optimal or near-optimal solution. These algorithms have broadly demonstrated its robustness solving problems of diverse characteristics. Actually, they have been used to estimate the parameters of various models as the least squares optimisation model (Lee, Park, & Chang, 2006; Wu et al., 2017; L. Yang, Chen, Rytter, Zhao, & Yang, 2019) to make accurate predictions of diverse problems.

In our research, we implemented a genetic algorithm designed specifically to estimate the optimal weights of a logistic regression model. The next sub-sections give a detailed explanation on both methodologies.

2.1. Logistic Regression

Logistic regression (Cox & Snell, 1989) is used for solving problems whose output variable is qualitative. Therefore, it has great applicability when the goal is to predict the appearance or not of a certain failure. The model establishes a probability of belonging to a class as a linear logistic expression (eq. 1).

$$p_i = \frac{1}{1 + e^{-wx_i}} \tag{1}$$

Where p_i, the probability of occurrence of a success of interest, is a function of x_i, the vector of explanatory variables, and w, their respective weights. The subscript i refers to each observation of the sample; $i = 1, \dots, N$. The model response is symmetrical as shown in equations (2- 4).

$$P(y = 1|x = x_i) = p_i \; ; \; P(y = 0|x = x_i) = 1 - p_i \tag{2}$$

$$1 - p_i = \frac{e^{-wx_i}}{1 + e^{-wx_i}} = \frac{1}{1 + e^{wx_i}} \tag{3}$$

$$p_i(x_i) = 1 - p_i(-x_i) \tag{4}$$

Weights, which are the parameters to be determined, are common for all instances *i*. They are usually estimated by maximising the log-likelihood function (eq. 7). The aim of the model is the assignment of a high probability of having the characteristic of interest to the observations with $y_i = 1$, and a low probability to those with $y_i = 0$. This function is obtained from the probability of a response, $y_i(0,1)$, (eq. 5).

$$P(y_i) = {p_i}^{y_i}(1 - p_i)^{1-y_i} \tag{5}$$

Assuming independence between observations, the likelihood function is obtained by (eq. 6).

$$l = P(y_1, \dots, y_N) = \prod_{i=1}^{N} {p_i}^{y_i}(1 - p_i)^{1-y_i} \tag{6}$$

Then, the use of the logarithm enables the conversion of the multiplication into a sum in order to make the resolution of the model easier. Finally, $ln\left(\frac{p_i}{1-p_i}\right) = wx_i$ and equation (4) are introduced in the function (eq. 6) which makes the form of the final log-likelihood function (eq. 7).

$$Log(l) = \sum_{i=1}^{N} y_i wx_i - log(1 + e^{wx_i}) \tag{7}$$

Once the weights have been estimated, the prediction of a new observation can be done by substituting its explanatory variables in equation (1). The obtained probability together with a pre-established risk threshold will determine the sample class (eq. 8). Although the threshold value is usually set to 0.5, it might be modified depending on the requirements of the problem.

$$y_i = \begin{cases} 0 \; if \; p_i \leq 0.5 \\ 1 \; if \; p_i > 0.5 \end{cases} \tag{8}$$

2.2. Genetic algorithms

Genetic algorithm (Holland, 1992) is a based on population metaheuristic which emulates Darwin's Theory of Evolution. This theory defends a process called natural selection; species that are better adapted to the environment are more likely to survive over time.

The designed GA seeks the optimal value of a group of real parameters. Although individuals are commonly codified as binary chromosomes, in this study each gene is a float between -1 and 1 representing explanatory variable's weights. Figure 1 shows the main steps of the applied GA: (i) selection of parent individuals from the population; (ii) crossover and mutation operators applied to these selected individuals; (iii) and the replacement on the population of the two new individuals.

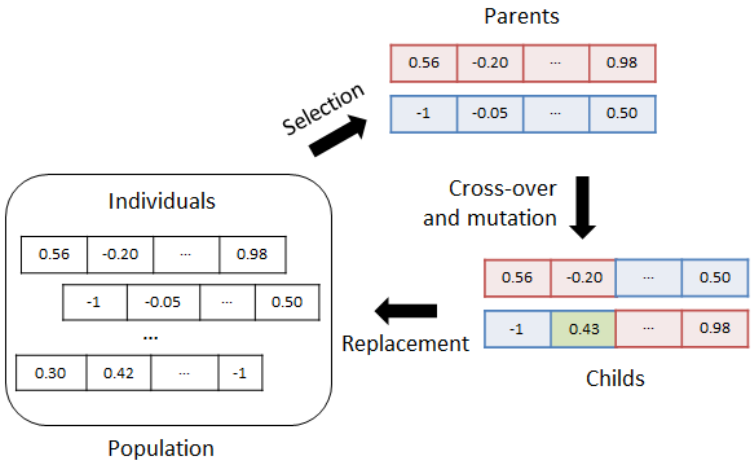

Fig 1 *A visual illustration of the designed genetic algorithm (figure created with MSOffice)*

Firstly, the algorithm starts with a set of individuals which is called population. Each individual is a solution to the problem and it is defined as a list of floats. The number of elements of the list must be equal to the number of explanatory variables plus one, the intercept. The population is randomly initialised, and its size is a control parameter which is obtained on the calibration phase.

Secondly, two individuals are randomly selected from the population because the elitist is introduced in the replacement phase. Moreover, if both parent individuals were equal, a new random individual would be generated to extend the search space.

Thirdly, the population is probabilistically modified by crossover and mutation operators. One-point crossover is applied because, in this problem, individuals are not too long. Regarding mutation, one weight of the individual is replaced by a new random one.

Finally, in order to maintain the best individuals of the population, a tournament replacement has been used. The two worst individuals, whose fitness functions are the lowest, of a group of individuals previously selected are replaced by the two new ones. The algorithm stops after a pre-fixed number of iterations.

3. Data description

We take data from a Spanish city. The analysed network has a total of 3,020km of sewer pipes. The historical database is composed of seven consecutive years (from 2012 to 2018), and it includes 3,917 failures. In this work, the output variable represents an incidence in the pipe, which requires some intervention in it. The incidents embrace both blockages and pipe breakages.

As previously mentioned, it is difficult to know in advance which factors are the most influential in the appearance of failures in sewer networks. Therefore, it is decided to add to the study every factor for which reliable data is available. The explanatory variables (x_i) together with the output variable (y) are listed in table 1. A number is assigned to each category of categorical variables starting by 0. Additionally, the percentage of network's length represented by each category (for categorical variables) or by ranges (for numerical variables) is shown in figure 2.

Table 1 *Description of variables*

	Variable	Type	Mean	Std	Min	Max
x_1	Material	Categorical	2.593	1.1E+00	0	4
x_2	Network type	Categorical	0.897	3.0E-01	0	1
x_3	Soil type	Categorical	1.133	5.1E-01	0	2
x_4	Water type	Categorical	1.904	3.0E-01	0	2
x_5	Section type	Categorical	3.813	6.8E-01	0	4
x_6	Diameter (mm)	Numerical	513.483	5.0E+02	75	6500
x_7	Length (m)	Numerical	24.752	2.2E+01	0.3	2523
x_8	Age (years)	Numerical	27.282	1.8E+01	-1	118
x_9	Exp. Sulphide	Numerical	0.066	1.1E-01	0	0.4
x_{10}	Previous failures	Numerical	0.033	2.3E-01	0	9
y	Pipe failure	Numerical	0.004	6.6E-02	0	1

Four materials compose 97% of the network length: reinforced concrete (RC), concrete (CON), polyvinyl chloride (PVC) and vitrified clay (VC) (figure 2-a). Most pipelines are made of CON (68.7%) followed by RC (15.9%). The rest of the materials are joined in one only group named '*others*'. There are two network types: the main collectors and the secondary network. The latter represents more than 85% of the network. The variable *soil type* differs between pipes under land, roadway or sidewalk. This can influence the loads they support. The variable *water type* informs about the origin of the water transported by each pipe which can be rainwater, sewage or both. As it can be seen in figure 2-d, most pipes conduct a combination of both, so this is a combined sewer system.

Unlike supply network pipes, which are usually circular, wastewater pipes' sections can have various shapes (figure 2-e). The diameter of non-circular sections, like oval or rectangular, is

estimated by equation (9), being h the height and w the width. Figure (2-f) shows that most of the network's pipes have an estimated diameter between 0.25 and 0.50 metres.

$$D_{eq} = \sqrt{\frac{4}{\pi} hw} \tag{9}$$

More than 96% of the network is composed by segments of a length lower than 100 metres. Both variables, *Diameter* and *Length*, have a much greater range than the rest as it can be appreciated in table 1. Therefore, they have been logarithmic transformed. The average age of the network is 27 years, being 14.9% of the network very new (less than 10 years). The exposure to hydrogen sulphide (*Exp. Sulphide*) takes into account the deterioration that wastewater can provoke on cementitious pipes, as internal corrosion. This variable is estimated based on the slope of the pipe segment.

The last figure (2-j) exhibits that more than 97% of the network has not suffered any prior failure, which implies that the database is totally unbalanced.

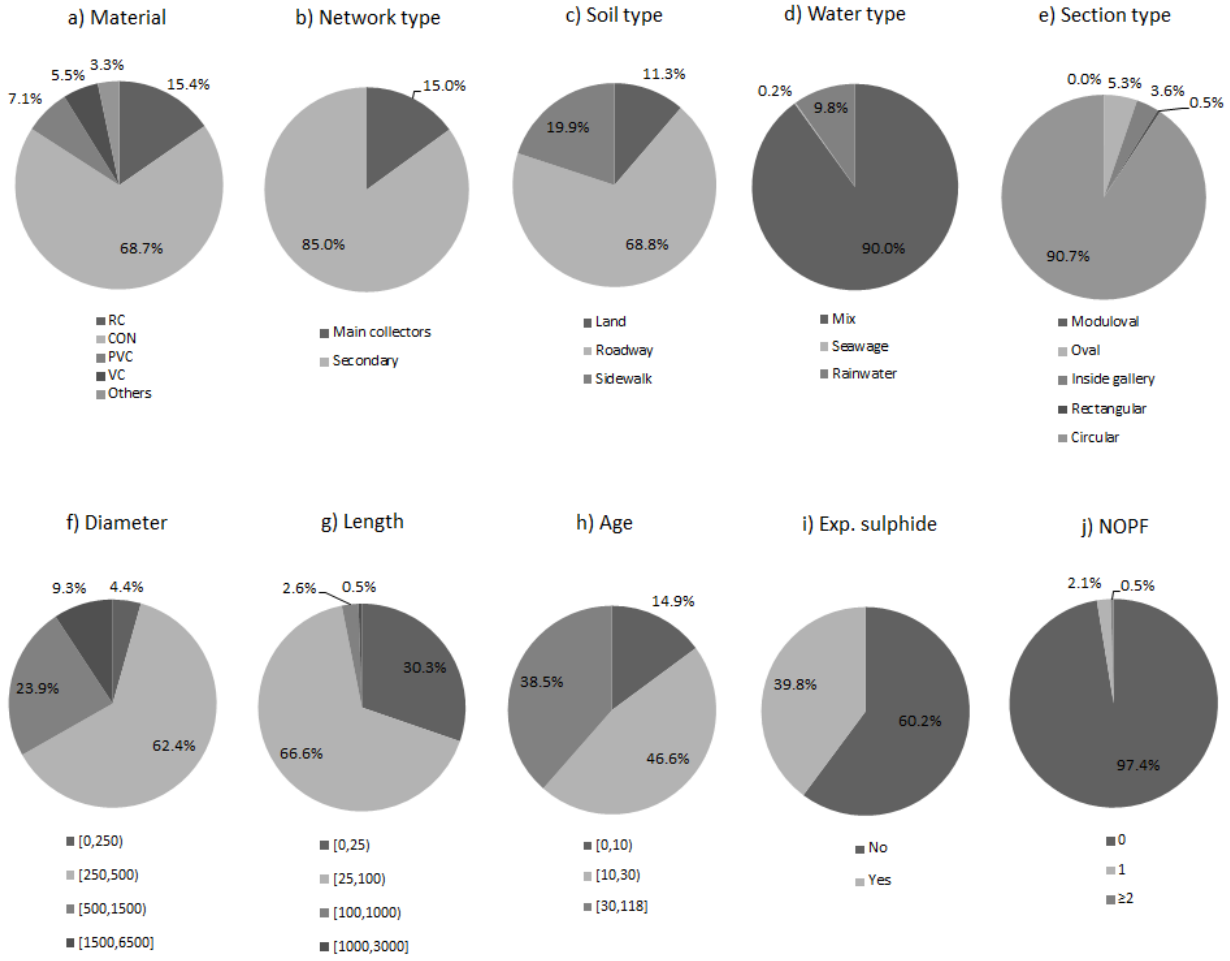

Fig 2 *Percentage of the network's length represented by each category of categorical variables, and by ranges of numerical variables (figure created with MSOffice)*

Figure 3 depicts the total number of failures, blockages and breakages of pipes, recorded in the available database. Despite the last year data, an increasing tendency over the years can be appreciated.

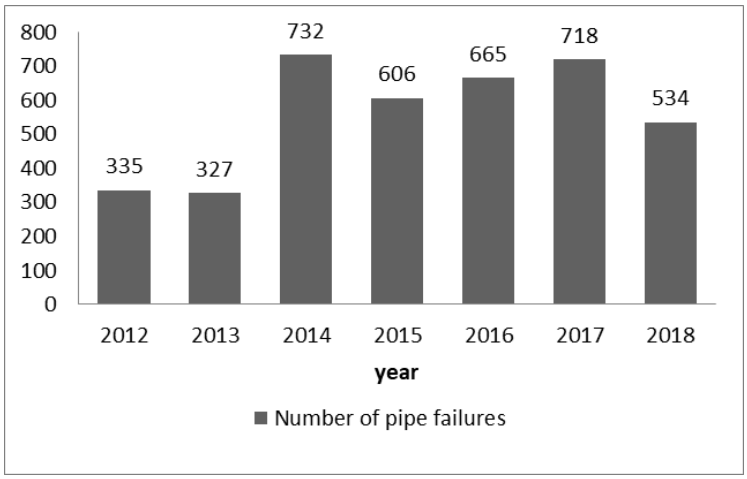

Fig 3 *Total recorded failures per year (figure created with MSOffice)*

4. Results and discussion

One of the main goals of a good classifier is to have good generalisation capabilities. For this purpose, not only should training data be representative, but overfitting must also be avoided (Flach, 2012). Five years from the historical database are used to train the model, which seems to be reasonably representative. Consequently, the test data, which assesses the performance of the estimated model, is composed by the remaining two years.

Data has a majority class of pipes which do not suffer any failure. However, it is more interesting to make the right predictions for those pipes which do fail, the minority class. For this reason, an under-sampling technique is implemented for the training set, so the algorithm can learn to make right predictions of both classes. Furthermore, missing data has been filled with the median of the variable.

To measure the methodology performance, two aspects should be considered: the GA convergence and the generalisation capabilities of the LR model. The former implies that the likelihood function of the best individual must increase over time which can be assessed by visualising this evolution. The latter, that the final estimated model must make right predictions. The confusion matrix and the receiver operational curve (ROC) are specific metrics to measure the performance of the binary classifier.

4.1. Calibration of the algorithms

Firstly, GA control parameters are calibrated together with the data scaling. A total of 24 simulations of 1,000 iterations (table 2) are carried out varying the population size (*Pop size*) and, the crossover and mutation probabilities (*CXPB* and *MUTPB*). Moreover, two different rescaling processes are applied to the input data: normalisation (eq. 10) and standardisation (eq. 11). The importance of data scaling lies in the diversity of variables' unit of measurements and the fact that weights are established in a range between -1 and 1.

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{10}$$

$$x_i = \frac{x_i - x_{mean}}{x_{std}} \tag{11}$$

On the one hand, normalisation reduces or extends the values of a variable to certain limits with respect to its minimum and maximum. If there was noise, this would be augmented with this transformation. On the other hand, standardisation converts data to a range of -1 to 1, making the new data to have a mean equal to 0 and a variance equal to 1.

Table 2 *Algorithm's calibration*

Sim.	1	2	3	4	5	6	7	8	9	10	11	12
MUTPB	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.5	0.5	0.5
CXPB	0.8	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.6	0.5	0.5	0.5
Pop size	100	200	1000	100	200	1000	100	200	1000	100	200	1000
Scaling	Stand	Stand	Stand	Stand	Stand	Stand	Stand	Stand	Stand	Stand	Stand	Stand
OF	-3188	-3225	-3330	-3183	-3211	-3290	-3187	-3195	-3377	-3213	-3257	-3338
Acc.	0.685	0.628	0.547	0.679	0.671	0.573	0.684	0.698	0.567	0.754	0.585	0.645
Recall	0.651	0.752	0.755	0.717	0.690	0.799	0.680	0.638	0.736	0.585	0.748	0.693
Sim.	13	14	15	16	17	18	19	20	21	22	23	24
MUTPB	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.5	0.5	0.5
CXPB	0.8	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.6	0.5	0.5	0.5
Pop size	100	200	1000	100	200	1000	100	200	1000	100	200	1000
Scaling	Norm	Norm	Norm	Norm	Norm	Norm	Norm	Norm	Norm	Norm	Norm	Norm
OF	-3375	-3425	-3442	-3369	-3416	-3478	-3421	-3403	-3421	-3412	-3393	-3457
Acc.	0.558	0.606	0.566	0.495	0.613	0.557	0.606	0.536	0.708	0.481	0.651	0.612
Recall	0.766	0.694	0.721	0.847	0.706	0.686	0.676	0.741	0.610	0.798	0.682	0.625

Looking at the objective function's values (OF) from table 2, it can be assumed that better results are obtained when GA control parameters are: a population size of 100 individuals, a high crossover probability of 0.7 and a mutation probability of 0.3. OF represents the log-likelihood function (eq. 7). To compare standardisation and normalisation, the value of OF is not relevant because each one generates different values of input variables. So, in this case, accuracy and recall of the test data are analysed.

Accuracy measures the percentage of well-classified instances, both 0 and 1. Despite being an important metric to quantify the quality of a classifier, for such data with unbalanced classes this can lead to confusion. For this reason, it is also essential to study the recall, which is the percentage of well-predicted instances from class 1 ($y_i = 1$). It can be appreciated that standardised data reaches more compensated values of accuracy and recall. Hence, this transformation seems to be more suitable.

4.2. Final results and quality metrics

Once control parameters are set, a new series of simulations is implemented in order to check the algorithm performance. Ten different simulations have been carried out, all of them with 10,000 iterations of the genetic algorithm. Table 3 shows the mean and the standard deviation of these simulations and the best attained solution. Control parameters are: a population of 100 individuals and a crossover and mutation probabilities of 0.7 and 0.3, respectively. Regarding data scaling, they have been standardised. According to test data results, it can be concluded that the algorithm generalises and the training data is representative.

Table 3 *Fitness function and quality metrics of the final solution*

	Training data				Test data		
	OF	Acc.	Recall	AUC	Acc.	Recall	AUC
Best sol.	-3154.08	0.669	0.681	0.738	0.674	0.690	0.765
Mean	-3154.23	0.668	0.683	0.738	0.671	0.690	0.765
Std dv	1.0E-04	9.2E-13	3.0E-11	5.6E-16	1.7E-11	1.3E-12	8.0E-14

The 67.4% of pipes are well-predicted and the rate of possible prevented failure with a threshold of 0.5 is 0.69. It can be noted that accuracy and recall are again really compensated.

Figure 4 shows the evolution of the log-likelihood function (OF) of the population's best individual in a simulation. It can be observed that it increases over time and, after approximately 6,000 iterations, it becomes stable. With this figure, the convergence of the GA is demonstrated.

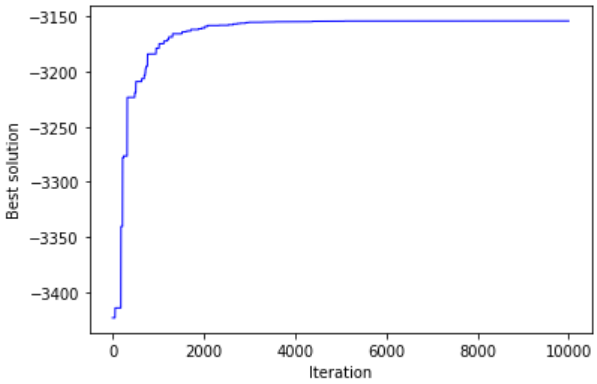

Fig 4 *Evolution of the best solution's likelihood function - 10,000 iterations (figure created with Python 3.7)*

The model estimated by the best solution achieves to predict a great number of pipe failures. In fact, replacing 3% of pipe segments, more than 25% of unexpected pipe failures could have been prevented in years 2017 and 2018 (test data). These are the segments whose probability is higher than 0.75 (see figure 5).

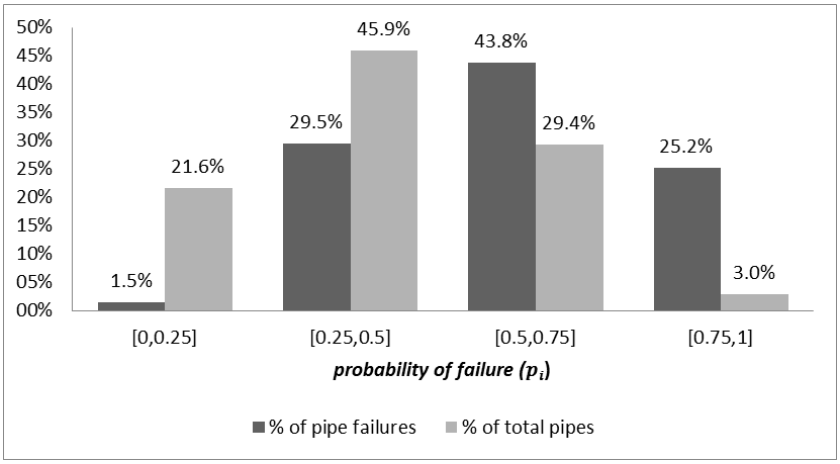

Fig 5 *Percentage of prevented failures and percentage of total pipes for different ranges of* $\boldsymbol{p_i}$ *(figure created with MSOffice)*

AUC is another quality metric which represents the ability of a classifier to avoid erroneous predictions (Fawcett, 2006). In figure 4, the ROC curve can be seen and its corresponding AUC for the test data of the best obtained solution of table 3. A perfect classifier would have an AUC of 1 while a random classifier would have AUC equal to 0.5. Our classifier demonstrates good capabilities of predicting if a pipe will or will not suffer a failure based on this case study. Every simulation reaches AUCs over 0.760 for the test data.

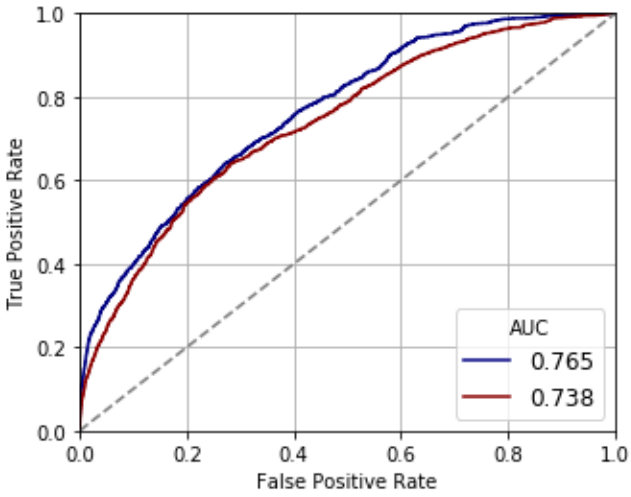

Fig 6 *ROC curves of the test and training data from the simulation of the best attained solution (figure created with Python 3.7)*

Finally, table 4 shows the optimised weights which inform us that the most influential variables causing failures are *diameter*, *water type* and *pipe age*. On the basis of weight sign analysis, it is concluded that smaller diameter pipes are more prone to fail. However, the weight of the variable *type of network* (-0.12) states that there is a greater risk of failure for pipes within the principal network which usually presents bigger diameters.

Table 4 *Logistic regression's weights obtained with the GA*

Variable	Weight	Best sol.	Mean	Std dv
Intercept	w_0	-0.45	-0.45	2.6E-10
Mat	w_1	0.18	0.18	2.2E-08
N_type	w_2	-0.12	-0.12	1.4E-08
Location	w_3	0.15	0.15	8.4E-10
Wat_type	w_4	0.37	0.38	1.9E-09
Sec_type	w_5	0.00	0.00	2.2E-08
log(DIA)	w_6	-0.70	-0.70	4.4E-08
log(LEN)	w_7	0.08	0.08	3.7E-09
Age	w_8	0.36	0.36	5.8E-10
Exp_sulf	w_9	0.09	0.09	1.7E-09
NOPF	w_{10}	0.22	0.22	4.1E-09

5. Conclusions

This study attains satisfactorily the proposed objective which was to predict pipe failures in sewer systems using a logistic regression model. Logistic regression was successfully used as a machine-learning classifier. The model was trained with reliable data from a high-dimensional network and the results were accurate. In this work, the unknown parameters of the model were estimated by a real-coded genetic algorithm. First of all, raw data was processed and, then, GA control parameters and data scaling were calibrated by several simulations. The importance of data pre-processing and the correct calibration of the algorithms was entirely verified. Once the model is trained, a probability of occurrence of the success of interest is assigned to each pipeline. It enables to follow different criteria depending on the available budget or the system's requirements. For instance, inspecting or replacing only those pipes whose probability was above a threshold. Moreover, results are very self-explanatory by non-statistics-experts, which facilitate its implementation in real world systems.

Final results demonstrate the algorithm convergence and the efficacy of the logistic regression model. 25.2% of incidents could have been prevented by inspecting 3% of the pipelines. It would suppose a great reduction of unexpected failures. Therefore, service quality would increase and the institution in charge could save significant costs.

A direct application of the developed methodology might be the optimisation of pipe's visual inspections schedules and maintenance tasks. Additionally, it allows obtaining information about the network conditions and the factors which play the most significant roles in the appearance of pipe's defects. On the contrary to previously suggested by (Savic et al., 2006), the weights of the logistic model inform that pipe diameter is the most influential variable and it has an inverse correlation with failure. Moreover, these pipes are more difficult to inspect directly by humans. Therefore, it would be interesting to prioritise the visual inspections of small pipes instead of the bigger ones. The type of transported water inside pipes is also a relevant variable. Pipes transporting combined water, which represent most of the network, have a higher risk of failure. As expected, older pipes are more prone to fail. However, it would be interesting to independently analyse which materials are more vulnerable to deterioration over time.

Future lines of research could combine the proposed methodology with the processing of images obtained from visual inspections. Definitively, it would improve the model accuracy.

Acknowledgements

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

6. References

Anbari, M. J., Tabesh, M., & Roozbahani, A. (2017). Risk assessment model to prioritize sewer pipes inspection in wastewater collection networks. *Journal of Environmental Management*, *190*, 91–101. https://doi.org/10.1016/j.jenvman.2016.12.052

Bailey, J., Keedwell, E., Djordjevic, S., Kapelan, Z., Burton, C., & Harris, E. (2015). Predictive risk modelling of real-world wastewater network incidents. *Procedia Engineering*, *119*(1), 1288–1298. https://doi.org/10.1016/j.proeng.2015.08.949

Cox, D. R., & Snell, E. J. (1989). *Analysis of Binary Data* (2nd ed.). London: Chapman and Hall Ltd.

de Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J. F. (2017). Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems with Applications*, *69*, 62–73. https://doi.org/10.1016/j.eswa.2016.08.014

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Flach, P. (2012). *Machine learning - The Art and Science of Algorithms that Make Sense of Data* (1st ed.). Cambridge: Cambridge University Press.

Halfawy, M. R., & Hengmeechai, J. (2014). Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine. *Automation in Construction*, *38*, 1–13. https://doi.org/10.1016/j.autcon.2013.10.012

Hassan, S. I., Dang, L. M., Mehmood, I., Im, S., Choi, C., Kang, J., … Moon, H. (2019). Underground sewer pipe condition assessment based on convolutional neural networks. *Automation in Construction*, *106*(April), 102849. https://doi.org/10.1016/j.autcon.2019.102849

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems* (2nd Editio). Michigan: MIT Press. Retrieved from https://ieeexplore.ieee.org/servlet/opac?bknumber=6267401

Khan, Z., Zayed, T., & Moselhi, O. (2009). Structural Condition Assessment of Sewer Pipelines. *Journal of Performance of Constructed Facilities*, *24*(2), 170–179. https://doi.org/10.1061/(asce)cf.1943-5509.0000081

Kleiner, Y., & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: physically based models. *Urban Water*, *3*(3), 151–164. https://doi.org/http://dx.doi.org/10.1016/S1462-0758(01)00033-4

Kleiner, Y., & Rajani, B. (2012). Comparison of four models to rank failure likelihood of individual pipes. *Journal of Hydroinformatics*, *14*(3), 659–681. https://doi.org/10.2166/hydro.2011.029

Kuliczkowska, E. (2016). Risk of structural failure in concrete sewers due to internal corrosion. *Engineering Failure Analysis*, *66*, 110–119. https://doi.org/10.1016/j.engfailanal.2016.04.026

Lee, Y. H., Park, S. K., & Chang, D. E. (2006). Parameter estimation using the genetic algorithm and its impact on quantitative precipitation forecast. *Annales Geophysicae*, *24*(12), 3185–3189. https://doi.org/10.5194/angeo-24-3185-2006

Li, D., Cong, A., & Guo, S. (2019). Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification. *Automation in Construction*, *101*(January), 199–208. https://doi.org/10.1016/j.autcon.2019.01.017

Mashford, J., Marlow, D., Tran, D., & May, R. (2011). Prediction of sewer condition grade using support vector machines. *Journal of Computing in Civil Engineering*, *25*(4), 283–290.

Savic, D., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S., & Saul, A. (2006). Modelling sewer failure by evolutionary computing. *Proceedings of the Institution of Civil Engineers - Water Management*, *159*(2), 111–118. https://doi.org/10.1680/wama.2006.159.2.111

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

Ugarelli, R., Kristensen, S. M., Røstum, J., Sægrov, S., & Di Federico, V. (2009). Statistical analysis and definition of blockages-prediction formulae for the wastewater network of Oslo by evolutionary computing. *Water Science and Technology*, *59*(8), 1457–1470. https://doi.org/10.2166/wst.2009.152

Wu, R., Painumkal, J. T., Volk, J. M., Liu, S., Louis, S. J., Tyler, S., … Harris, F. C. (2017). Parameter estimation of nonlinear nitrate prediction model using genetic algorithm. *2017 IEEE Congress on Evolutionary Computation, CEC 2017 - Proceedings*, (3), 1893–1899. https://doi.org/10.1109/CEC.2017.7969532

Yamijala, S., Guikema, S. D., & Brumbelow, K. (2009). Statistical models for the analysis of water distribution system pipe break data. *Reliability Engineering and System Safety*, *94*(2), 282–293. https://doi.org/10.1016/j.ress.2008.03.011

Yang, M. Der, & Su, T. C. (2009). Segmenting ideal morphologies of sewer pipe defects on CCTV images for automated diagnosis. *Expert Systems with Applications*, *36*(2 PART 2), 3562–3573. https://doi.org/10.1016/j.eswa.2008.02.006

Yang, L., Chen, G., Rytter, N. G. M., Zhao, J., & Yang, D. (2019). A genetic algorithm-based grey-box model for ship fuel consumption prediction towards sustainable shipping. *Annals of Operations Research*. https://doi.org/10.1007/s10479-019-03183-5

Yang, M. D., & Su, T. C. (2008). Automated diagnosis of sewer pipe defects based on machine

learning approaches. *Expert Systems with Applications*, *35*(3), 1327–1337. https://doi.org/10.1016/j.eswa.2007.08.013

Younis, R., & Knight, M. A. (2010). A probability model for investigating the trend of structural deterioration of wastewater pipelines. *Tunnelling and Underground Space Technology*, *25*(6), 670–680. https://doi.org/10.1016/j.tust.2010.05.007