Machine learning to improve understanding of sewer pipe failures

Ehsan Kazemi, Ph.D.[1], Will Shepherd, Ph.D.[1*], Simon Tait, Ph.D.[1]

[1]*Department of Civil and Structural Engineering, University of Sheffield, Sheffield, UK*
**Corresponding author email: w.shepherd@sheffield.ac.uk*

Keywords: Machine Learning; Self-Organising Maps; Random Forests; Blockage; Flooding.

Highlights

- Machine Learning shows strong promise for interpreting linked pipe asset and failure data
- Random Forests are able to predict if a pipe is at risk of causing a failure with high accuracy
- Through predicting failure probability, pipes at risk can be identified for proactive inspection

Introduction

Historical data collected from a sewer network covering a town with a population of about 50k, with 180 km of predominantly combined sewers is analysed to understand potential causes of observed failure incidents, such as blockage and flooding. For this purpose, Machine Learning (ML) models are developed to identify the relationships between different elements of the system, and then estimate the risk of incidents through quantifying the identified relationships. This risk relationship can be used to plan pro-active inspection to identify developing issues and so ensure more focussed risk-based maintenance.

Data and pre-processing

The data includes an asset inventory comprising 8128 pipes that gives characteristics such as pipe diameter, length and gradient. A separate dataset of recorded incidents on the network, includes 1409 blockages (SB), 1380 flooding incidents (SF) and 538 incidents of other types (SO), totalling 3327 over 7 years between 2015 and 2022. The two datasets were joined together in ArcGIS to create an 'incident' input data for ML analysis. The joining was performed by linking recorded incidents (which are associated to street addresses) to the nearest asset through an ArcMap proximity search. Besides, a 'non-incident' dataset was prepared to be combined with the incident dataset and used in the ML analysis. This dataset includes 3327 pipes without any incidents during the 7 year period. The non-incident samples are necessary for quantifying probability of incidents in the supervised ML. Apart from the attributes of the network provided in the asset database, several additional parameters were calculated or extracted from other databases such as the UK Meteorological Office Rain Radar Data (Met Office, 2003). The extra parameters included Soil Water Index (Copernicus Service information 2022), maximum rainfall intensity on the incident date, total precipitation volume on the day of the incident and the previous 2, 3, 4 and 7 days, gradient of steepest upstream pipe, gradient of shallowest upstream pipe, largest upstream pipe diameter, smallest upstream pipe diameter, change in gradient, change in diameter.

Approach

Firstly, a qualitative analysis using an unsupervised ML technique, Self-Organising Map (SOM), is carried out to identify plausible linkages between network characteristics and the incident categories to identify the more important parameters relating to the risk of incidents. Then, a supervised ML risk model is developed based on Random Forests to quantify the relationships.
Two tests were performed as follows. I) Output parameter was defined whether any type of incident occurred on each pipe. Therefore, it has two classes, incident (shown by 'F') or non-incident (shown by 'N'). There are 3327 incidents and 3327 non-incidents, thus 6654 samples in the data. II) the output parameter is defined by specific types of incidents. Thus, the parameter has four classes, non-incident (N), flooding (SF), blockage (SB) and other failure types (SO). The size of the classes in the data is different. ML techniques often have difficulties with imbalanced datasets see Kazemi et al.

(2022a), on the performance of Random Forests in dealing with water system imbalanced datasets. Therefore, the majority classes were randomly down-sampled so that all the classes had 538 samples, leading to a dataset of 2152 samples.

SOM preliminary analysis

SOM is a type of unsupervised Artificial Neural Networks (ANN) for data clustering. All the parameters are fed into the model as input parameters and the linkages between them are qualitatively and visually investigated. Various combinations of parameters introduced in the 'Data and pre-processing' section were tested and the parameters listed in Table 1 showed to be most relevant (the definition of changes in gradient and diameter of the pipes are depicted in Figure 1). Figure 2 presents the SOM developed for the best input parameter combination. On the lattices, each hexagonal cell (neuron) represents a group of samples; colours show the value of the variables (red: high, blue: low); and each cell in the same position on different maps corresponds to the same group of observations. In addition to the maps of parameters, an additional lattice called unified distance matrix, or 'U-matrix' is provided which shows the strength of the clusters. The Incident Type lattice on the right is post labelled to the SOMs to clearly show how the incident types are grouped.

Table 1. Parameters used in the SOM in Figure 2.

Parameter	Definition
gradient	Pipe gradient
usLargSteepGradChange	Change in pipe gradient vs largest steepest pipe at upstream
diameter	Pipe diameter (mm)
usBigDiamChange	Diameter change from the upstream pipe with the largest diameter
length	Pipe length (m)
populationDensity	People per (Km^2)
swi	Soil Water Index for the day that the incident was reported (%)
maxIntensity0	Maximum rainfall intensity for the day that the incident was reported (mm/hr)
totalPrecip0	Total precipitation for the day that the incident was reported (mm)
dataQual	Data quality – a score of how much of length, diameter and material is missing

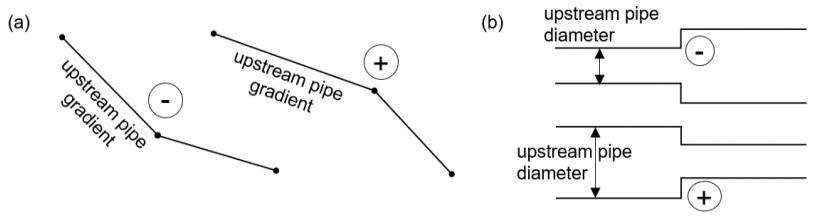

Figure 1. Change of gradient (a) and diameter (b) from an upstream pipe (definition of positive and negative values).

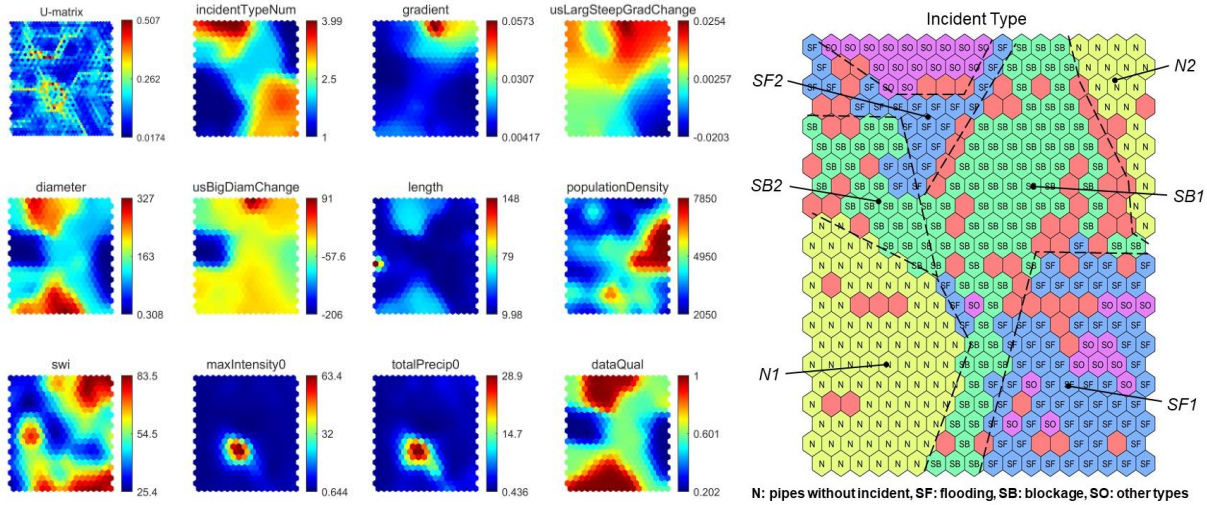

Figure 2. SOM developed to classify incident types. Parameters are defined in Table 1. *IncidentTypeNum* is an integer from 1 to 4 (1 = N, 2 = SB, 3 = SF, 4 = SO).

The SOM presented in Figure 2 has mainly classified the incidents into well-defined areas. Although direct correlations are not seen between most of the parameters and the incident type, the latter has partial linkages with some of the parameters, particularly the ones related to the pipe gradient. The major findings are that i) larger cluster of non-incidents (*N1*) correlates with low pipe gradients where change in gradient from upstream pipes is close to zero; ii) pipes with flooding are mainly linked to very low gradients and negative values of gradient change (*SF1*) (see Figure 1 for the definition of a negative gradient change); and iii) blockages are mostly correlated with positive gradient changes (*SB1*), i.e., transition from shallower to steeper pipes, and most of these are when *usBigDiamChange* is positive, i.e., there is a decrease in pipe diameter. A smaller group of blockages (*SB2*) are linked to negative values of *usBigDiamChange*, where the pipe diameter is small.

Risk model

Random Forests were chosen for the development of the risk model because previous studies showed that this method outperformed other supervised techniques like Support Vector Machine and Boosted Decision Trees learning algorithms (e.g., LogitBoost, RobustBoost and RUSBoost) in data-driven prediction of probability of water systems failures (Kazemi et al., 2022a and 2022b, for drinking water distribution systems and railway drainage systems, respectively).
As discussed in the 'Approach' section, two tests were performed, namely Tests 1 and 2, with the target being predicting incidents from an incident/non-incident dataset (F and N) and predicting incident types from a dataset comprising different types of incidents (SF, SB, SO and N). For each test, the parameters identified by SOM to be important (see Table 1) were used as input; and the data was split into two subsets of 85% and 15%, where the former was used to train the model and the latter was employed to verify the performance of the trained model in predicting unseen data.

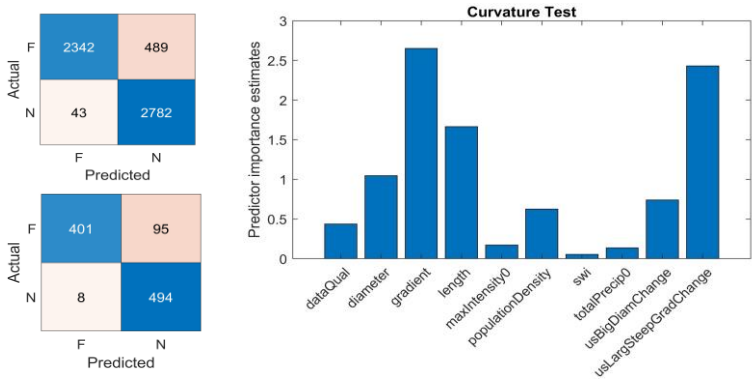

Figure 3. Confusion matrices for training and validation sets (top left and bottom left, respectively) and predictor importance estimates (right) for Test 1.

Figure 3 presents the confusion matrices for the training and validation sets as well as the result of a curvature test to identify the importance of the predictors for Test 1. The confusion matrices show that the model was capable of predicting incidents with a very high accuracy. In predicting unseen data, 401 of 496 incidents and 494 of 502 non-incidents were predicted correctly. This has led to a True Positive Rate (*TPR*) of 0.809, a True Negative Rate (*TNR*) of 0.984, an Accuracy (*ACC*) of 0.897 and an F_1-score of 0.906 for the validation set. *TPR* and *TNR* represent the rate of correct predictions of incidents and non-incidents, respectively; and *ACC* and F_1 represent the overall accuracy of the model. Table 2 presents the performance metrics for both training and validation sets.
According to the curvature test for Test 1 (Figure 3-right), the most important parameters were pipe gradient and change in pipe gradient vs largest steepest pipe at upstream. Then, length, diameter, change in diameter and population density were important; and the rainfall parameters (total precipitation, maximum intensity and SWI) were the least important factors in predicting incidents.
Figure 4 shows the confusion matrices for Test 2, i.e., when the target is incident type (SF, SB, SO or N). The model predicted non-incidents (N) with a high accuracy (*ACC* of 0.899 for unseen data), but the prediction of the incident type (SF, SB and SO) is significantly lower, particularly in terms of *TPR* and F_1. All the accuracy measures for the training and validation sets for Test 2 are shown in Table 2.

The results of Tests 1 and 2 suggest that, for the present data, the model performs quite well in predicting whether a pipe will fail, but its accuracy is limited in predicting the type of the failure.

Table 2. Performance metrics for Tests 1 and 2 for the training and validation sets.

Metrics	Test 1: incident/no incident		Test 2: incident types					
	Incident (class F)		No incident (class N)		Flooding (class SF)		Blockage (class SB)	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
TPR	0.827	0.809	0.960	0.909	0.707	0.391	0.698	0.403
TNR	0.985	0.984	0.931	0.896	0.947	0.820	0.942	0.847
ACC	0.906	0.897	0.938	0.899	0.887	0.713	0.881	0.740
F_1–score	0.898	0.906	0.886	0.822	0.758	0.403	0.746	0.427

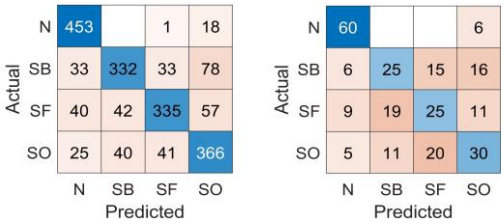

Figure 4. Confusion matrices for training and validation sets (left and right, respectively) for Test 2.

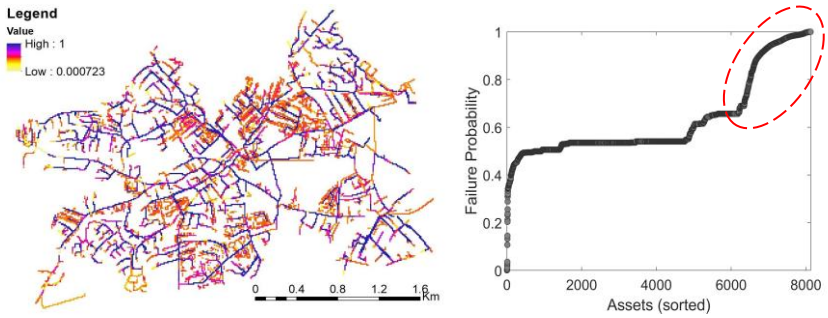

Figure 5. Relative probabilities of incidents calculated for all the pipes in the system.

The Random Forests model trained in Test 1 was then employed to predict relative incident probability of all the pipes in the network. Figure 5 shows the result on a GIS map as well as a ranking of the pipes based on the predicted probability. The ranking shows that only about 20% of the pipes (shown by dashed line) pose a high risk (>0.7) compared to the entire system, suggesting that targeting these for inspection should provide the greatest return for risk-based maintenance.

Conclusions

SOM was used to identify the most relevant parameters related to the failure of sewer pipes, and these were then employed in a supervised ML model to predict probability of failures. The ML model performance in predicting specific failures, like flooding and blockage, was not very high; but it was able to predict whether a pipe is at risk of any type of failure incident or not with a very high accuracy. The trained model was then employed to predict relative probability of failures for the whole network, and this was useful for risk ranking of the pipes for targeting them for inspection.

Acknowledgements

This work is supported by the EU's H2020 research and innovation programme grant no. 101008626 and the UK's Engineering and Physical Sciences Research Council grant EP/S016813/1.

References

Copernicus Service information (2022): https://land.copernicus.eu/global/products/swi (accessed on 01 August 2022).

Kazemi E., Kyritsakas G., Husband S., Flavell K., Speight V., Boxall J. (2022a). Predicting iron exceedance risk in drinking water distribution systems using machine learning. 14th International Conference on Hydroinformatics, Bucharest, Romania.

Kazemi E., Wu Y., Nichols A., Tait S. and Raja J. (2022b). Machine learning for data driven management of UK railway drainage infrastructure. 39th IAHR World Congress. Granada, Spain.

Met Office (2003): Met Office Rain Radar Data from the NIMROD System. NCAS British Atmospheric Data Centre, 2020. http://catalogue.ceda.ac.uk/uuid/82adec1f896af6169112d09cc1174499 (accessed on 01 August 2022).