ENAC / PROJET DE MASTER 2020-2021
SECTION DE GÉNIE CIVIL

Smart Management of Underground Infrastructure

Generic Model to predict the Likelihood of Failure for water pipes

Author : Mohammadhossein Esmaeelzadeh

Supervisors in company : Dr. Karim Claudio [1] / Mr. Thomas Van Becelaere [2]

Supervisor in EPFL : Prof. Dr. Dimitrios Lignos [3]

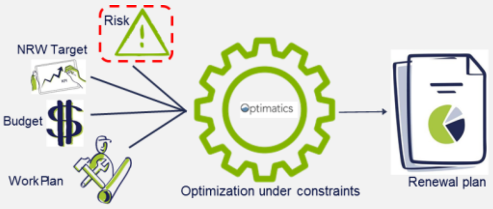
[1] SUEZ Environments – East Asia / [2] SUEZ Smart Solutions / [3] Resilient Steel Structures Laboratory (RESSLAB)

Introduction and Context

Problem statement: In the lack of high-quality data with little or no information about the past bursts in a water network, is it possible to use the historical bursts of similar pipes and learn from these past bursts to predict the Likelihood of Failure?

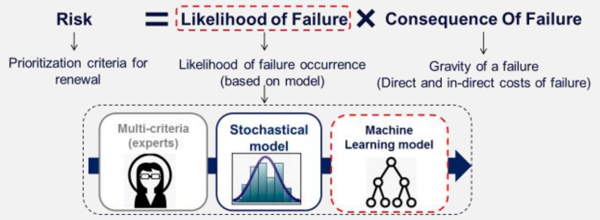
Remarks

- Burst of the pipes could be due to many different parameters and not only the age.
- For many small networks and networks in developing countries, the historical bursts records are whether shallow or not reliable.
- For networks with basic/rich data, SUEZ and many other competitors propose Statistical and Machine Learning solutions, but there is a gap for contracts with poor data.
- Some competitors such as FRACTA recently has proposed new solutions for poor data networks. SUEZ has currently no solution for this market.



- In this project, we have attempted to fill this gap in SUEZ and have proceeded the following steps:

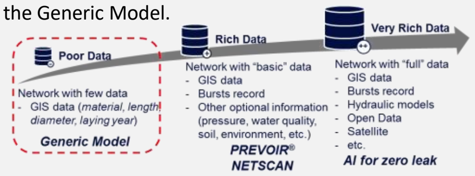
1. Created the SUEZ worldwide database for pipes and bursts.
2. Enriched the SUEZ database with open-source data to introduce the context of the pipes.
3. Design a Machine Learning model to predict the Likelihood of Failure for a new network without the historical bursts records by training it on similar pipes in the SUEZ database.
4. Automate this machine learning pipeline.


Risk
=
Likelihood of Failure
X
Consequence Of Failure
Prioritization criteria for renewal
Likelihood of failure occurrence (based on model)
Gravity of a failure (Direct and in-direct costs of failure)
Multi-criteria (experts)
Stochastical model
Machine Learning model

Renewal planning in SUEZ

By having the risk associated with the failure of each pipe, the available budget, Non-Revenue Water target and work plan constraint, we formulate an optimization problem to get the renewal plan.

In this project, we predict the Likelihoodhood of Failure, which is needed for this optimization problem to calculate the risk. Depending on the quality of data, there are different methods for predicting the LoF. Figure below, demonstrates these solutions in SUEZ. This project is a strating point for the Generic Model.



Data collection

Pipes and bursts record data:

The data of pipes and bursts of more than 1000 cities with 175000 km of pipes in 10 countries are validated and cleaned. This data was validated, cleaned and summarized to find the anomalies and patterns. The results of these summaries were discussed with domain experts. We have prepared a database of pipes and their historical bursts.

Open-source data:

These open-source data are from satellite images. We have enriched the database of pipes and bursts with these data. We have added information about the soil, climate, vegetation, urbanism, population, topography etc. to introduce the context of each pipe.

Remark. These data were collected previously, but we have validated and cleaned them.

Methodology

We have designed the Generic Model to predict the LoF for water pipes, by using the enriched database for pipes and bursts.

Generic Model

For a new network, the procedure is as follow:

Step 1: Enrich the new network data with open-source data to introduce the context of the pipes.

Step 2: Preprocessing, including getting the required columns from the SUEZ database, handling missing values, one hot-encoding for categorical columns, normalizing data according to their distribution in SUEZ database and dimension reduction with Principal Component Analysis.

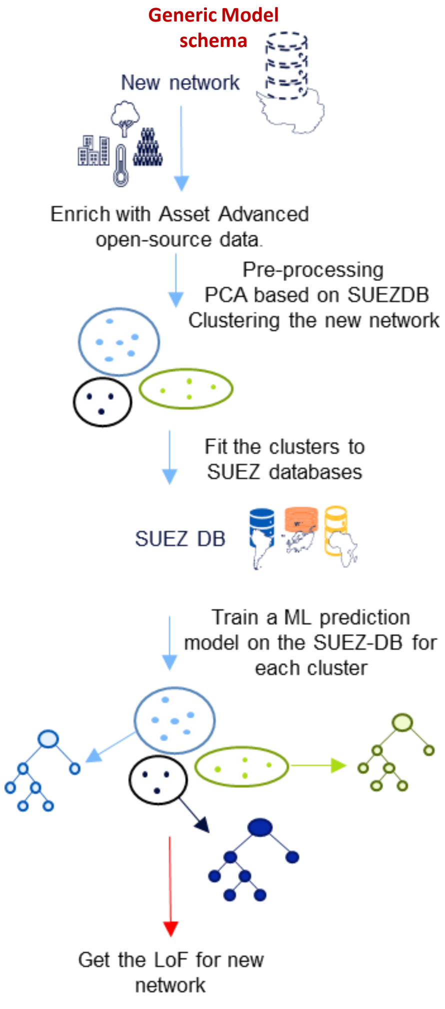
Step 3: Clustering the new network with k-means method to group pipes with similar features.

Step 4: For each cluster, fit the pipes in SUEZ database that are closest to the centroid, to find 20 times of similar pipes in the corresponding cluster.

Step 5: Train an ML classifier for each cluster for the pipes of the SUEZ database in each cluster.

Step 6: Predict the LoF for the new network based on the trained classifier in each cluster.

Remark. The Generic Model and the data are on CODAi, a cloud-base machine learning platform belonging to SUEZ hosted by Microsoft AzureML. This model is optimized to have a low computation cost. This makes using it easier and tuning its hyperparameter more convenient.


Generic Model schema
New network
Enrich with Asset Advanced open-source data.
Pre-processing
PCA based on SUEZDB
Clustering the new network
Fit the clusters to SUEZ databases
SUEZ DB
Train a ML prediction model on the SUEZ-DB for each cluster
Get the LoF for new network

Results

KPIs for LoF models

ROC Curve: plot FPR vs. TPR for different thresholds.

Failure Avoidance Curve: Sort the pipes from the highest LoF to lowest LoF and plot the accumulated length percentage vs. accumulated bursts percentage.

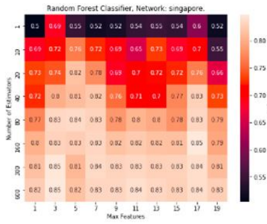
Choice of KPI

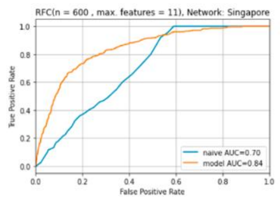
AUC of ROC curve was chosen for the following reasons:

- Demonstrate the results for different thresholds.
- A conventional method to evaluate the performance of the Machine Learning and LoF models.
- Effect of the length requires further study: long pipes are more prone to have a burst, but misclassification of a very long pipe has a high penalty. Thus, we need to decrease the length contribution to LoF.

How were the results produced?

We have performed a customized cross-validation. Since most of the pipes are from France (135000 km out of 175000km), we have created a validation set by choosing all the network outside France and one network in France. We have treated each of the validation networks as a new network. So, we take out that network from the training set and evaluate the performance of the model with each of these networks. We have used the Area Under Curve of ROC as a KPI for the performance of the model. In the figure below, we see a heatmap to optimize two parameters of the classifier. Also, we see a ROC curve for a tuned model with the results with a naïve analysis, merely based on age..





Conclusion

We have started the project by creating an enriched SUEZ database. This database is used as a descriptive tool, to discover the past events, learn from these events to understand why these bursts happened and utilize this learning for predicting future events to improve decision making. Then, we have designed the Generic Model from enriching data to preprocessing and getting the Likelihood of Failure. Normalizing the data is required for dimension reduction and clustering. The PCA is performed for reducing the computation costs and handling the colinear features. The clustering and fitting the clusters is a kind of soft clustering, so if a pipe is close enough to multiple clusters, it could be used for all of them. Clustering reduced the computation time significantly and learning from pipes with similar features is more relevant.

A few percentages of improvement in leaks in the scale of the city has a significant environmental and economical effect. The first result of the model is by choosing the worst 10% of pipes we can identify about 30 to more than 50% of the bursts.

The proposed model is a starting point and will be further improved in the future.

An advantage of the Generic Model is to produce results faster than a multi-criteria analysis. Also, we don't need a local expert to perform the analysis. On the other hand, the Generic Model is learning from data. So, for a new network, the results should be used carefully as there might be a specific network with different reasons for bursts which the model is not trained for.