Leak detection using Random Forest and pressure simulation

L. Aymon, J. Decaix, F. Carrino, P-A. Mudry, E. Mugellini, O. A. Khaled, R. Baltensperger
University of Applied Sciences and Arts Western Switzerland
lucien@aymon.me, {jean.decaix, pierre-andre.mudry}@hevs.ch
{francesco.carrino, elena.mugellini, omar.aboukhaled, richard.baltensperger}@hefr.ch

***Abstract*—Water is a scarce resource which is becoming increasingly inaccessible. It is therefore necessary, in most parts of the world, to capture, transport and allocate it efficiently and thoughtfully. The implementation of monitored water distribution networks is often expensive. The purpose of this project is therefore to monitor leakage and consumption in a non-pressurized agricultural irrigation system using only inexpensive and easily installed pressure sensors. We modeled the water network to automatically simulate a leak randomly through the network. These simulated pressures serve as a dataset to train, test and validate a Random Forest algorithm that detects the leaks. Through pressure measures, the model can locate the junction closest to the leak with an accuracy of 96.24%. This approach therefore allows leaks detection in a water distribution system without the use of expensive flow sensors.**

***Index Terms*—Machine Learning, Random Forest, Water distribution network, EPANET**

I. Introduction

Monitoring water distribution systems (WDS) allows water consumption to be tracked and makes the detection of anomalies, such as leaks, possible. Knowing the flows gives the possibility of evaluating the behaviour of the networks. However, the installation of flow sensors in irrigation networks is time-consuming and requires a significant economic investment.

The flow rate in an irrigation consumption points (faucets) depends on many physical factors, such as its altitude, the geometry of the valve and the state of the network, which is represented in our study by pressures at different locations.

The purpose of this project is to estimate flows in the water distribution system based on pressures measured on the consumption points and, based on this information, to detect possible leaks in the network. Differently from flow sensors installed on pipes, pressure sensors are quickly and inexpensively installed above ground, on faucets.

This work is based on data coming from pressure sensors installed on the agricultural irrigation network of Bagnes (VS) in Switzerland.

We modeled the WDS of Bagnes using a simulator calibrated on collected real data, which has the advantage of reducing the needs of actual sensors on the real network. Then, we use the simulator to generate a dataset with leaks, or not, placed at random positions in the network. Finally, we trained a machine learning model (Random Forest) on this dataset to automatically detect and locate leaks.

II. Related work

Machine learning is used in many WDS-related studies. It can optimize the allocation schedules, determine the water needs by soils and plants, predict soil yield based on the amount of water allocated, predict futures consumption, detect leaks and estimate their flows [1].

To our knowledge, no publication used machine learning for the detection of the consumption based on the pressures values in the network. On the other hand, many publications described leaks detection projects based on machine learning and, many of these, where based on simulated flows. Reference [2] used Support Vector Machine (SVM) achieving an accuracy of 76.8%. (Note that this value depends on the number of junctions in the network). On a simplified network and for detection based on leakage groups rather than individual valve, logistic regression reached an accuracy score higher than 90% [3]. These results show that the accuracy achieved by machine learning to detect junction or junction group near a leak is accurate enough to be transposed to consumption detection (consumption flows can be seen as leaks with a high flow rate).

Machine learning algorithms are also widely used to predict water consumption. With advanced deep learning models, such as the Artificial Neuro-Genetic Networks in [4], a prediction of 93% of the observed water demand variability with a standard error prediction of 12.63% for a one-day forecast and with actual data in limited quantities.

Even if deep learning for leak detection seems a promising approach, we propose the use of a Random Forest model. Random Forest has the advantage of requiring less input data than approaches based on Deep Learning (even if this is not a limitation for a simulated network, it will be a relevant problem when we will use our system to handle real data). In addition, Random Forest has already been used for leak detection on different scenarios [5]. Finally, the Bagnes network is not pressurized and, to our knowledge, Random Forest has not been used to detect leaks in this type of scenario.

III. Methodology and first results

We modeled part of the WDS of Bagnes on the EPANET water distribution network management software [6]. A script has been developed in Python to simulate the pressure in the network, with successive openings of the faucets. At this point, one or zero leak are present in each simulation, each

simulation will represent an entry in the input dataset of the machine learning model. The main goal is to know which faucets or groups of faucets consume water and therefore locate a leak. The simulation allows all pipelines to be monitored and eliminates the need to manually open valves or add leaks to the physical network.

A. Leaks simulation and detection

The simulation of leaks is done by adding a consumption node on a pipe randomly selected in the network and at a random position over the length of the pipe. We repeated this process many times to generate the dataset. The Random Forest classifier learns from the training set to locate the position of the leak. The number of labels (i.e., the closest junction to the leak) is equal to the number of distribution points and therefore it could be quite high (e.g., there are around 150 distribution points for the modeled part of the Bagnes network). The learning curve shows that we need the number of pipes times 2000 to properly train and validate our Random Forest model.

Grid Search is used to perform hyperparameter optimization of Random Forest. To evaluate the achieved results, we used k-Fold cross validation ($k = 5$). The hyperparameters explored by the algorithm are: the maximum depth of the tree, the minimum number of samples required to split an internal node, the number of features to consider when looking for the best split and the maximum depth of the tree. We have set the limits for these parameters based on experience.

B. First results

The Grid Search algorithm provided the following results: Bootsrap is *True*, min. samples split is *3*, max. features is *3*, and max. depth is *None*.

Fig.1 shows the learning curve for Random Forest classification trained on the Bagnes WDS with a dataset of 276'000 samples (each one of them with one or zero simulated leak in a randomly selected pipe). This high number of simulations is required by the high variability introduced by randomizing the position of the leaks on the pipes. Comparing the accuracy on the validation and training set, it seems that the algorithm starts converging after 200'000 samples to reach an accuracy of 96.24%, and a F1-score of 95.59%.

C. Conclusion and Future works

We presented an approach to generate a dataset of leaks of a WDS using a simulator. The generated data has been used to train a Random Forest classifier to detect the leaks on the network. The results look promising but a deeper investigation is still required to validate the results and to test other scenarios (such as networks with multiple simultaneous leaks). Despite Random Forest gives us good results, other ML algorithms will still be tested, in order to improve the accuracy. As the results of several studies show, Artificial Neural Networks [7] and more generally deep learning algorithms, can provide leakage estimates with interesting precision for our project and therefore they will be included in our analysis. In addition,

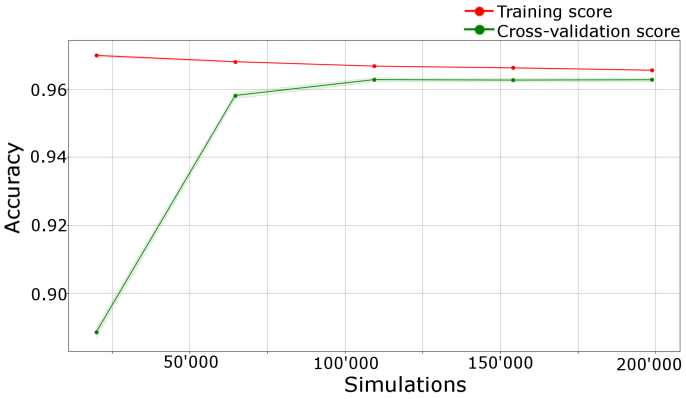

Fig. 1. Learning curve for Random Forest

we plan to use Transfer Learning to transfer the knowledge acquired with the simulated data to real data, which will be collected in limited sections of the Bagnes network.

Due to the complexity of the WDS of this study, consumption groups will be created and, in a second step, the size of these groups will be reduced until they come as close as possible to a granularity of a single consumption point (allowing, for instance, to bill individual users for their water consumption). We also plan to improve the error estimation by considering the distance between the predicted leak and the actual one.

Since it is possible to compare water consumption to a leak, it will be possible to predict future uses of the network. These forecasts will be based on the consumption history, on external factors (e.g., weather forecasts) and field measurements (e.g., soil humidity). The knowledge of consumption trends in the short and medium term will makes it possible to anticipate water shortages and suggest effective countermeasures.

IV. Acknowledgment

The research program iNUIT is part of the large-scale thematic research programs of the HES-SO. The authors gratefully acknowledge funding from HES-SO during the period of 2013-2020.

References

[1] A. Goldstein et al., "Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge," Precision Agriculture, vol. 19, pp. 421–444, 01.07.2018.

[2] J. Mashford and D. De Silva and S. Burn and D. Marney, "Leak detection in simulated water pipe networks using SVM," Applied Artificial Intelligence, vol. 26, pp. 429–444, 2012.

[3] G. Gupta, "Monitoring Water Distribution Network using Machine Learning," 2017.

[4] R. G. Perea and E. C. Poyato and P. Montesinos and J. A. R. Diaz, "Irrigation Demand Forecasting Using Artificial Neuro-Genetic Networks," Water Resources Management, vol. 29, pp. 5551–5567, 2015.

[5] Z. Chen and X. Xu and X.Wang and X. Zhong, "DEStech Transactions on Engineering and Technology Research," 2018.

[6] EPANET, Application for Modeling Drinking Water Distribution Systems
`https://www.epa.gov/water-research/epanet`

[7] M. Shinozuka and J. Liang and M. Q. Feng, "Use of Supervisory Control and Data Acquisition for Damage Location of Water Delivery Systems," Journal of Engineering Mechanics, vol. 131, pp. 225–230, 2005.