Robles-Velasco, Alicia; Muñuzuri, Jesús; Onieva, Luis; Rodríguez-Palero, María

Article

Trends and applications of machine learning in water supply networks management

Journal of Industrial Engineering and Management (JIEM)

Provided in Cooperation with:
The School of Industrial, Aerospace and Audiovisual Engineering of Terrassa (ESEIAAT), Universitat Politècnica de Catalunya (UPC)

Suggested Citation: Robles-Velasco, Alicia; Muñuzuri, Jesús; Onieva, Luis; Rodríguez-Palero, María (2021) : Trends and applications of machine learning in water supply networks management, Journal of Industrial Engineering and Management (JIEM), ISSN 2013-0953, OmniaScience, Barcelona, Vol. 14, Iss. 1, pp. 45-54, https://doi.org/10.3926/jiem.3280

This Version is available at:
https://hdl.handle.net/10419/261740

OmniaScience

Journal of Industrial Engineering and Management
JIEM, 2021 – 14(1): 45-54 – Online ISSN: 2013-0953 – Print ISSN: 2013-8423
https://doi.org/10.3926/jiem.3280

Trends and Applications of Machine Learning in Water Supply Networks Management

Alicia Robles-Velasco, Jesús Muñuzuri, Luis Onieva, María Rodríguez-Palero

Dpto. Organización Industrial y Gestión de Empresas II. Universidad de Sevilla. (Spain)

arobles2@us.es, munuzuri@us.es, onieva@us.es, mariarodriguez@us.es

Received: July 2020
Accepted: November 2020

Abstract:

Purpose: This study describes the trends and applications of machine learning systems in the management of water supply networks. Machine learning is a field in constant development, and it has a great potential and capability to attain improvements in real industries. The recent tendency of data storage by companies that manage the water supply networks have created a range of possibilities to apply machine learning. One particular case is the prediction of pipe failures based on historical data, which can help to optimally plan the renovation and maintenance tasks. The objective of this work is to define the stages and main characteristics of machine learning systems, focusing on supervised learning methods. Additionally, singularities that are usually found in data from water supply networks are highlighted.

Design/methodology/approach: For this purpose, thirteen studies which contain real cases from around the world are discussed. From the data processing to the model validation, a tour of the methods used in each study is carried out. Moreover, the trendiest models are briefly defined together with the mechanisms that best suit their performance.

Findings: As a result of the study, it was found that the imbalanced class problem is typical of data from water supply networks where only a small percentage of pipes fail. Consequently, it is recommended to use sampling methods to train classifiers, however, it is not necessary if we are training a regression system. Additionally, scaling and transformation of variables has generally a positive impact on the model's performance. Currently, cross-validation is almost a requirement to obtain reliable and representative results. This technique is employed in most revised studies to train and validate their models.

Originality/value: The use of machine learning systems to predict pipe failures in water supply networks is still a developing field. This study tries to define the advantages and disadvantages of different methods to process data from water supply networks, as well as to train and validate the models.

Keywords: machine learning, supervised learning, water supply networks, pipe failures, predictive systems

To cite this article:

Robles-Velasco, A., Muñuzuri, J., Onieva, L., & Rodríguez-Palero, M. (2021). Trends and Applications of Machine Learning in Water Supply Networks Management. *Journal of Industrial Engineering and Management*, 14(1), 45-54. https://doi.org/10.3926/jiem.3280

1. Introduction

Within the integral water cycle there are many possible machine learning applications to optimise processes and support decision-making. For instance, concerning sewer networks, there are currently emerging techniques to treat and classify images from inside the pipes to detect leaks and anomalies (Li, Cong & Guo, 2019). Reviewing thousands of images in real time is a difficult task for humans; however, machine learning systems can do it in a few seconds. Therefore, the integration of these systems as a support tool can drastically decrease the number of unexpected incidents and thus the emergency response time.

Regarding water supply networks, there is a global tendency of management companies of this service that is the introduction of machine learning to predict pipe failures and breakages. In the case of Spain, its water supply infrastructure is composed of more than 256,984 km of pipes and 39% of them are over 30 years old (AEAS, 2016). There is an evident aging of the infrastructures, and the occurrence of unexpected leaks and breaks remains a problem that concerns every management company. As previously said, this is not only a national problem, but a global one. The solution lies in leveraging data and applying machine learning algorithms to reduce the number of unexpected pipe failures.

This paper presents an introduction of machine learning and its applications to water supply network management. Specifically, the main stages of its implementation and its most critical aspects are reviewed. Additionally, the mechanisms employed to address these critical aspects are described by thirteen researches that use supervised machine learning systems to predict pipe failures on water supply networks. These papers have been chosen because most of them present real cases studies from around the world and include reliable data. Therefore, we can analyse the singularities that are usually found in data from water networks. Furthermore, the selected researches allow revising the most popular machine learning techniques in this field and they provide rich explanations of their applications to the case studies.

2. Machine Learning: Concepts and Stages

Machine learning is a field of artificial intelligence that gathers algorithms and techniques that allow creating systems abled to learn from experience. These systems must generalise behaviours and recognise patterns from data. There are three different machine learning systems: supervised learning systems, unsupervised learning systems and reinforced learning systems. Depending on the data nature and the output variable to be predicted, a type of them must be chosen.

Supervised learning requires labelled data, i.e., the output variable must be identified and available. If the output variable is a real value, regression methods are the most appropriate. And, classification systems are suitable when the output variable is a category or a class. In both cases, the final objective is predicting.

Unsupervised learning is used when there are not data labels, or they are not clearly identified. The main objective of these systems is to extract knowledge and discover hidden patterns in data. Clustering is the most representative unsupervised learning technique. Finally, reinforced learning systems interact with the environment, receiving feedback. Therefore, its performance improves over time. Autonomous vehicles are the most famous example of this kind of systems.

This study focuses on supervised learning applications as predictive systems in supply water networks. These techniques are the most common in this industry because of its easy integration with support decision system tools. Figure 1 shows the stages of supervised learning system's implementation. Firstly, data is divided into training and validation sets; secondly, the training data is used to estimate the parameters that define the machine learning model; and then the performance of the model is measured through certain quality metrics over the validation set. These three first stages, data processing, training, and validation of the model, are discussed in the subsections below. The last stage, prediction, is basically the use of the system to forecast future behaviours based on new data.

In order to show the main problems and solutions that can arise from the implementation of a supervised learning system with data from water supply networks, thirteen studies are analysed. Table 1 presents the references together with the models they apply, the predicted output variables and some details about the real case studies they use to evaluate the performance of these models. The acronym N.M. means Non-Mentioned in the article.

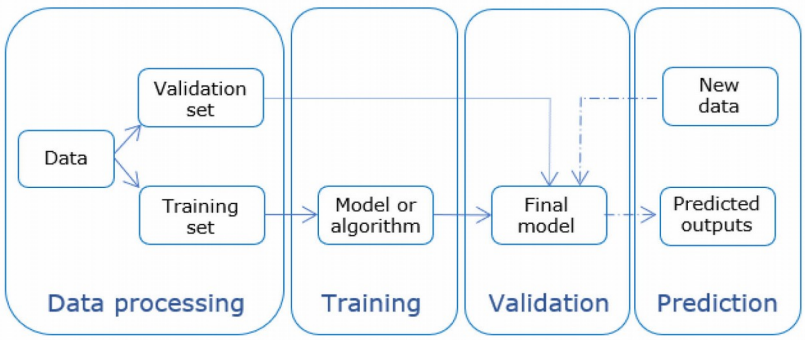

Figure 1. Implementation stages of a supervised machine learning system

On the one hand, networks have been stratified in three categories according to their length as: (1) large, greater than 3,000km; (2) medium, between 1,000 and 3,000 km; and (3) small, lower than 1,000km. Moreover, the number of recorded pipe failures has been added as it is linked with the robustness of the conclusions and the scope of the study. On the other hand, the country of each case study has been included to demonstrate that these techniques are being applied all around the world. Concretely, the country which is producing more scientific literature about this topic is Canada. Therefore, four studies from this country has been included. In general, there are obstacles to access data in this industry because of two main reasons: firstly, companies privacy policies; and secondly, the lack of robust data storage plans. Nevertheless, the second one has experienced a notable improvement in the last decade. Nowadays, companies are more aware of the enormous potential of data and this is encouraging the arise of new and more elaborated predictive models in this field. For more information on the case studies, readers are invited to consult the references.

Finally, models and output variables are discussed on sections 2.2 because they deserve a more detailed explanation.

References		Model	Output	Real case study		
				Network length	No. pipe failures	Country
1	(Kabir, Tesfamariam & Sadiq, 2015).)	SM	Time to failure	Large	4949	Canada
2	(Sattar, Gharabaghi & McBean, 2016)	GP	Time to failure	Medium	9500	Canada
3	(Sattar, Ertuğrul, Gharabaghi, McBean & Cao, 2019)	ANN	Time to failure	Medium	9500	Canada
4	(Kutyłowska, 2018)	SVR; ANN	Failure rate	Small	88	Poland
5	(Shirzad, Tabesh & Farmani, 2014)	SVR; ANN	Failure rate	Small	686	Iran
6	(Birek, Petrovic & Boylan, 2014)	EFS	Pipe leakages	N.M.	N.M.	UK
7	(Almheiri, Meguid & Zayed, 2020)	ANN; SM; DT	Time to failure	Small	N.M.	Canada
8	(Royce, Seth & Henneman, 2014)	BBNs	Number of pipe failures in a zone	Large	3686	USA
9	(Robles-Velasco, Cortés, Muñuzuri & Onieva, 2020)	LR; SVC	Failure probability	Large	4393	Spain
10	(Tang, Parsons & Jude, 2019)	BBNs	Failure probability	N.M.	N.M.	UK
11	(Wang, Dong, Wang, Tang & Yao, 2013)	RankBoost	Risk index	Large	11603	China
12	(Winkler, Haltmeier, Kleidorfer, Rauch & Tscheikner-Gratl, 2018)	DT	Failure	Small	3743	Austria
13	(Tavakoli, Sharifara & Najafi, 2020)	RF	Failure	Large	N.M.	USA

Table 1. References, models, output variables and their case studies

2.1. Models: Characteristics and Applications

Previous studies have extensively implemented physical and statistical models to describe the water network behaviour and to analyse the pipes failures in order to find patterns and weaknesses. On the one hand, physical models try to determine pipes failures through the simulation of different external and internal conditions that have a negative effect in the pipe structure. Its main advantage is that they do not require large amount of data (Wilson, Filion & Moore, 2017). On the other hand, statistical models create statements based on observed data. These models are used with descriptive purposes which means, they do not predict any output variable. Its main limitation is that they cannot discover complex relationships between variables. In Scheidegger, Leitão and Scholten (2015), an extensive review of statistical models and its application to water supply networks can be found. It needs to be mentioned that statistical models are the base of many machine learning algorithms. Nevertheless, machine learning systems do require considerable amount of data, but they have a wider scope: They can be used both as descriptive models and as predictive ones and can discover complex hidden patterns.

There are references on the use of many different supervised learning models in the water network industry. In this study, eleven of the most common and trendy ones are selected (view Table 1). The main characteristics of these models and its acronyms are briefly defined below.

Artificial Neural Networks (ANNs) are famous because of its accuracy and talent to extract patterns from data (Kutyłowska, 2018; Sattar et al., 2019; Shirzad et al., 2014). These models try to emulate the functioning of the human brain where neurons are represented by nodes and the nerve impulses by a weighted sum of input variables. The learning process consist in the adjustment of its parameters, while the network structure does not usually vary. They have excellent generalisation capabilities. However, these models do not allow interpreting the role of each variable in the process of prediction and need large amount of data to be trained. Support vector machines can be used for regression (SVR) and classification (SVC) purposes. This method maps the explanatory variables through non-linear structures into a high dimensional space and then, the hyperplane that optimally adjusts to the data or separates the classes is generated (Kutyłowska, 2018; Robles-Velasco, Cortés, Muñuzuri & Onieva, 2020; Shirzad et al., 2014). Both ANNs and SVMs are informally known as 'black-box' systems. In contrast, survival models (SMs) and logistic regression (LR) provide a precise interpretability of results but they have more limitations to extract patterns from data. While SMs predict the life or time to failure of the instances (Kabir et al., 2015), LR is typically used for classification tasks (Robles-Velasco et al., 2020).

Genetic Programming (GP) is an evolutionary methodology that uses an iterative process to find the equation that best fits the relationship between several previously stated variables (Sattat et al. 2016). This method gives a detailed description of the system behaviour. However, if the equation is too complex, conclusion extraction is difficult, and the training process becomes computationally inefficient.

Fuzzy logic uses fuzzy sets and rule matrices to classify or categorise samples. This technique has been implemented in many water supplies studies to group pipes or regions of the networks according to their risk of failure (Al-Zahrani, Abo-Monasar & Sadiq, 2016; Islam, Sadiq, Rodriguez, Najjaran, Francisque & Hoorfar, 2013; Salehi, Jalili Ghazizadeh & Tabesh, 2018). Nevertheless, in all these studies the rules are generated based on expert opinions. Recently, a new application of fuzzy logic that include evolutionary algorithms to estimate the rules and parameters of the systems have appeared. They are referred as Evolutionary Fuzzy Systems (EFS) and are more independent and accurate that the traditional ones. The main advantage of EFS is the straightforward interpretability of results in the form of simple rules. As a disadvantage, its training is computationally expensive, and the design of these systems has a substantial dependency on the case study since many parameters must be decided in advance. In Birek et al. (2014), it is proposed an EFS together with a clustering strategy of data whose final goal is to predict leakages in water supply systems. It needs to be mentioned that this method has not been sufficiently explored nor applied in the water field yet.

Bayesian Belief Networks (BBNs) are graphical representations as direct acyclic graphs where nodes represent parameters and arcs the probabilistic relationship between them (Royce et al., 2014; Tang et al., 2019). They give a global vision of the relationship between every pair of variables. For this reason, it is convenient to include in the model all the available variables. This technique allows accomplishing a diagnostic and prognostic analysis.

Decision Trees (DT) are simple and computationally efficient methods that can be used with regression and classification purposes. The predictor space is stratified into a finite number of regions using splitting rules which are hierarchically combined into a tree (Winkler et al., 2018). Its major advantage lies in the direct visualisation of the relationship between variables which allows detecting the most vulnerable points of water networks. As a disadvantage, DT easily leads to overfitting of data. In Almheiri et al. (2020), it is suggested a boosting technique to reduce the prediction errors of single decision trees. Another option is to use Random Forests (RF) which combine a huge number of decision trees and aggregate their predictions (Wu & Baker, 2020). In Tavakoli, Sharifara and Najafi (2020), this technique is used to predict the pipe condition of sewer pipes in order to optimally plan the inspections according to the risk of failure of each area. Although this reference does not predict failures in water supply networks, it is included in the study because the use of variables and the processing of data is similar to the rest of revised studies. Moreover, the RF algorithm is exhaustive explained.

Finally, RankBoost is a boosting-type algorithm that makes bipartite ranking (Wang et al., 2013). The final demand of most companies is a ranking of the pipes according to their risk or probability of failure. Therefore, this kind of methods is really suited to face this problem.

Prior to the election of the model, the priority between the accuracy of results or the interpretability and the role of variables must be defined. ANNs and SVMs are recommended when the priority is the accuracy of results. If the objective is to analyse and interpret the results and the role of the variables, statistical models, decision trees or BBNs are greater alternatives.

2.2. Data Processing

Data processing may be the most important stage to construct a robust and accurate predictive system. Most data from water supply networks share similar characteristics, so they can be processed using the same techniques. As previous experience is helpful, Table 2 gathers the responses that the cited studies have given to different aspects related to data processing. The reference numbers correspond to the ones presented in Table 1.

Ref.	Missing values	Outliers	Feature selection	Scaling or transformation
1	N.M.	N.M.	Covariate selection process	Log
2	N.M.	Removed	Genetic Algorithm	None
3	N.M.	N.M.	Sensitivity analysis	None
4	Maintained	Maintained	None	None
5	N.M.	N.M.	Sensitivity analysis	Normalisation
6	N.M.	N.M	Experts criterion	Normalisation
7	N.M.	Removed	Availability	N.M.
8	Maintained	Maintained	Estimation per area	Standardisation
9	Median	Median	Experts criterion	Log and standardisation
10	Proxies	Removed	None	Discretisation
11	Removed	N.M.	Assessment by Sliding Thresholds	None
12	N.M.	N.M	Curation	None
13	Removed	Removed	Experts criterion	None

Table 2. Techniques applied in the data processing stage of each study

Both missing values and outliers are common in most databases and they are generally due to errors in data collection, or to some unusual circumstance. While the former are gaps of information, the latter are atypical values that a variable takes which are far from the main trend of the rest of data. Generally, it is recommended to eliminate the observations which contain these anomalies if they are not considered representative (Tang et al.,

2019; Wang et al., 2013). Nevertheless, it implies information losses, thus it is sometimes preferable to use the mean, the median or a proxy of the variable to fill or replace these values (Robles-Velasco et al., 2020). Other option is to use truncated distributions from available datasets to determine these data as it is done in the reference 2 from Table 2 (Sattar et al., 2016).

A high number of input variables might imply negative consequences as slowness in the training phase or difficulties in the results interpretation. Although this situation is not common with data from water networks whose variables are usually scarce, not all variables influence breakage. Therefore, it is recommended to seek the optimal set of variables based on certain quality metrics. Sometimes, they are chosen based on expert opinions. Nevertheless, a more technical option is to use some feature selection technique as filters or brute force. Filters are applied before the training of the model (Sattar et al., 2019; Shirzad et al., 2014) and they are based on statistical parameters or even graphics. 'Brute force' is the term used to define the process of training the model iteratively with different groups of variables until the most significant one is found (Kabir et al., 2015; Wang et al., 2013). This technique is more accurate, but it is also much more computationally expensive.

Finally, scaling and transformation of variables has more relation with the machine learning model, because some of them exhibit a high sensitivity to variable scale. Firstly, the normalisation of the variables (Eq. 1) has demonstrated to be useful for training ANNs. Secondly, the standardisation (Eq. 2) reduces the effect of outliers which are typical in databases. Finally, the logarithmic transformation is recommended if some variable extends into higher orders of magnitude, which it usually happens with the diameter or the length of pipes comparing with other variables such as the age or the pressures inside pipes. This transformation is especially useful to train statistical models. In (Winkler et al., 2018), it is stated that decision trees do not require data to be transformed before training.

$$X_i = (x_i - x_{min}) / (x_{max} - x_{min}) \quad (1)$$

$$x_i = (x_i - x_{mean}) / x_{std} \quad (2)$$

2.3. Training and Validation

Training and validation stages are strongly linked. In the training phase, the parameters that govern the model are estimated. The objective is to find the parameters that optimise some quality metric using a set of data, usually referred as training set. Most times, the same metrics are employed to train and validate the model. Table 3 includes information about the use of two techniques that appear in most studies. On the one hand, cross-validation is an iterative training-validation process that allows obtaining more accurate results and avoiding overfitting. It consists in dividing the data into several sets and training the model with a part of them, and then validates it with the rest. Figure 2 shows a diagram of a 3-fold cross-validation process. As can be seen in Table 3, almost all studies have employed this technique. In the studies 5 and 10, the dataset is divided into three groups: training, test and validation. Validation data do not participate in the training process and results are purely obtained from this unseen data. In this case, cross-validation is implemented using the training and test sets in order to estimate the parameters of the final model.

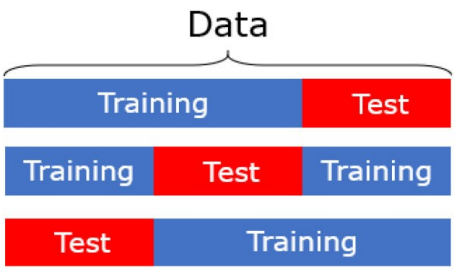

Figure 2. 3-fold cross-validation

Ref.	Cross-validation	Imbalanced class	Quality metrics
1	5-fold (80-20%)	Different consideration	Error measures
2	4-fold (75-25%)	Installation year bunch	Error measures and R^2
3	4-fold (75-25%)	Not considered a problem	Error measures and R^2
4	4-fold (75-25%)	N.M.	Error measures
5	Training (60%), test (20%) and validation (20%)	N.M.	Error measures and R^2
6	No. Training and test (80-20%)	N.M.	Error measures and R^2
7	5-fold (80-20%) and validation (10%)	N.M.	Error measures and R^2
8	No	Different consideration	R^2 and significance
9	5-fold (80-20%)	Under-sampling	Confusion matrix, ROC
10	Training (60%), test (20%) and validation (20%)	N.M.	Confusion matrix, ROC
11	5-fold (80-20%)	Under and over sampling	ROC
12	2-fold (50%-50%)	Under-sampling	Confusion matrix, ROC
13	No. Training and test (70-30%)	Bootstrap	Confusion matrix, ROC

Table 3. Aspects of the training and validation phases of each study

On the other hand, there is a necessity to deal with the imbalanced class problem, which is present in all databases of historical pipe failures from water supply networks. There are more pipes which do not fail than pipes which do. If the ratio exceeds 1:10, the learning task is considered as an imbalanced learning problem. This situation may imply negative repercussions on the behaviour of the model, especially if it is a classifier. However, some studies argue that the presence of imbalanced classes has not always worsen the performance of predictive models since it depends on the model and the data structure (Wang et al., 2013). Most of the presented classification studies address this problem by sampling the data (see Figure 3). This consists in eliminating samples (under-sampling) or generating new and artificial ones (over-sampling) in order to reduce the unbalance between the two classes in the dataset. Under-sampling has the disadvantage of losing valuable data while over-sampling can generate erroneous patterns so the training set is not representative. The choice of one technique or the other must be based on the number of recorded pipe failures in the dataset. Provided that the amount of recorded pipe failures is representative, it is preferable to use under-sampling. Meanwhile, over-sampling is the best option if the number of pipe failures in the dataset is scarce.

The studies that use regression models usually give a different treatment to pipes that fail and to the ones that do not fail, or simply do not mention this fact.

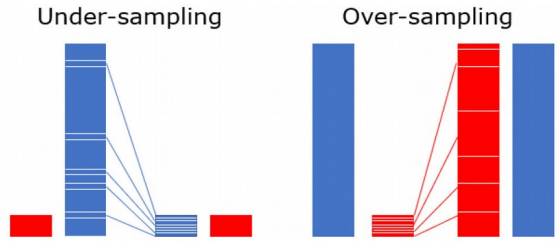

Figure 3. Under-sampling and over-sampling techniques

In general, the results of these studies must be interpreted by workers of water companies that are usually non-experts on machine learning. Therefore, quality metrics should be carefully chosen in order to faithfully represent the behaviours of the model and to make them easier to interpret. Quality metrics are numerical measures that represent the performance of a model and each model suggests a specific one.

Regressive methods are validated with error measurements as Mean Squared Error (MSE), Mean Absolute Error (MAE) or determination coefficient (R^2). Their formulas are presented in Equations 3, 4 and 5. These metrics quantify the differences or deviations between predicted and real system outputs. It is important to consider that not all metrics represent the same scale as variables. The relative error is perfect to compare different measures since it is a percentage (Kutyłowska, 2018).

$$MSE = \frac{\sum_{i=1}^{n}(y_{real} - y_{pred})^2}{n} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_{real} - y_{pred}|}{n} \tag{4}$$

$$R^2 = \frac{\sum_{i=1}^{n}(y_{pred} - \overline{y_{real}})^2}{\sum_{i=1}^{n}(y_{real} - \overline{y_{real}})^2} \tag{5}$$

The confusion matrix and the Receiver Operating Characteristic (ROC) curve are specific tools to evaluate classification models. On the one hand, the confusion matrix quantifies the number of correct and incorrect predictions for each class. It is an easy interpretable metric which allows extracting a lot of information. On the other hand, the ROC curve represents graphically the true positive rate against the false positive rate for different thresholds (see Figure 4). The Area Under the Curve (AUC) is a numerical measure between 0 and 1 that allows comparing different models. The closer to one is the AUC, the more accurate is the model.

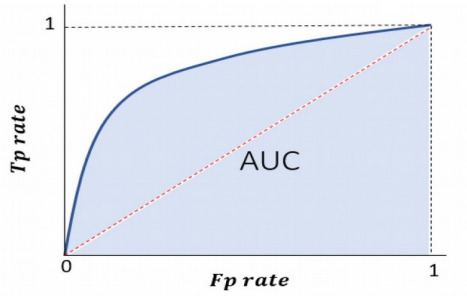

Figure 4. ROC curve

As can be seen in Table 3, the references 1-7 obtain a real value as output variable, while references 9-13 obtain a classification of the pipes. Reference 8 is a special case since it estimates all the variables per area. Although it is a classification system, its quality metrics are more typical of regression systems.

3. Conclusions

This paper presents an analysis of the application of machine learning techniques to the management of water supply networks. For this purpose, the main stages of the implementation of supervised learning models are reviewed, as well as some characteristics and critical aspects that appear when working with data from water supply networks. Moreover, the mechanisms used in thirteen studies to solve these difficulties are described.

Firstly, it is observed that there are many studies around the world that apply machine learning to enhance water supply networks management. Seven of the reviewed studies show medium or large networks with historical databases that exceed 3,500 pipe failures. Thus, the conclusions of these studies are representative.

Secondly, eleven different models and its applications, such as supervised learning systems, are briefly described in this study. Their main characteristics and the differences between them are highlighted, discovering the importance of initially establishing the aim of the system in order to choose the most suitable model. If accuracy of results prevails, it is recommended to use ANNs or SVMs. Nevertheless, if the objective is to analyse and interpret the results and the role of the variables, statistical models or BBNs are a better option. In the revised studies, several types of output variables have been found, as ranking, failure rate or time to failure of each pipe. Consequently, in a

real case application, this should be discussed with the company managing the water supply network before designing the machine learning system because it must be integrated with its decision-making tools.

Regarding data processing, it is important to mention the tendency of applying feature selection techniques instead of expert opinions. Additionally, scaling and transformation of variables have demonstrated to be positive.

Finally, this work wants to encourage the use of machine learning systems in the water network industry because of their independency and accuracy. Although many studies have recently emerged, there is still a gap in the real application of these techniques, and they have demonstrated to be useful and robust. Furthermore, it is noticed that the access to data is generally scarce. Therefore, water companies should consider enhancing their data collection system as well as facilitating researchers access to them.

In future research, the variables employed in each study and their influence on the pipe failures could be analysed. Plans for replacement and maintenance of pipes in water networks usually include supply and sewer pipes. Therefore, a similar analysis on machine learning models applied to sewer networks would complement this paper. Data from these networks usually include images, so image processing techniques should also be revised.

Acknowledgments

The authors wish to acknowledge to EMASESA, Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla, and to the Universidad de Sevilla (VI PPIT-US) because of their financial support through the Distinguished Cátedra del Agua.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

AEAS (2016). *XIV Estudio Nacional de Suministro de Agua Potable y Saneamiento en España*. http://www.aeas.es/servlet/mgc?pg=ListNews&ret=next&news_id=1249&areaCode=publicarea&newsCategory=Noticias

Al-Zahrani, M., Abo-Monasar, A., & Sadiq, R. (2016). Risk-based prioritization of water main failure using fuzzy synthetic evaluation technique. *Journal of Water Supply: Research and Technology - AQUA,* 65(2), 145-161. https://doi.org/10.2166/aqua.2015.051

Almheiri, Z., Meguid, M., & Zayed, T. (2020). Intelligent Approaches for Predicting Failure of Water Mains. *Journal of Pipeline Systems Engineering and Practice,* 11(4), 1-15. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000485

Birek, L., Petrovic, D., & Boylan, J. (2014). Water leakage forecasting: The application of a modified fuzzy evolving algorithm. *Applied Soft Computing Journal,* 14, 305-315. https://doi.org/10.1016/j.asoc.2013.05.021

Islam, M.S., Sadiq, R., Rodriguez, M.J., Najjaran, H., Francisque, A., & Hoorfar, M. (2013). Evaluating Water Quality Failure Potential in Water Distribution Systems: A Fuzzy-TOPSIS-OWA-based Methodology. *Water Resources Management,* 27(7), 2195-2216. https://doi.org/10.1007/s11269-013-0283-6

Kabir, G., Tesfamariam, S., & Sadiq, R. (2015). Predicting water main failures using Bayesian model averaging and survival modelling approach. *Reliability Engineering and System Safety,* 142, 498-514. https://doi.org/10.1016/j.ress.2015.06.011

Kutylowska, M. (2018). Forecasting failure rate of water pipes. *Water Science and Technology: Water Supply,* 19(1), 264-273. https://doi.org/10.2166/ws.2018.078

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, 199-208. https://doi.org/10.1016/j.autcon.2019.01.017

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliability Engineering and System Safety,* 196. https://doi.org/10.1016/j.ress.2019.106754

Royce, A.F., Seth, D.G., & Henneman, L. (2014). Bayesian Belief Networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering and System Safety,* 130, 1-11. https://doi.org/10.1016/j.ress.2014.04.024

Salehi, S., Jalili Ghazizadeh, M., & Tabesh, M. (2018). A comprehensive criteria-based multi-attribute decision-making model for rehabilitation of water distribution systems. *Structure and Infrastructure Engineering,* 14(6), 743-765. https://doi.org/10.1080/15732479.2017.1359633

Sattar, A.M.A., Ertuğrul, Ö.F., Gharabaghi, B., McBean, E.A., & Cao, J. (2019). Extreme learning machine model for water network management. *Neural Computing and Applications,* 31(1), 157-169. https://doi.org/10.1007/s00521-017-2987-7

Sattar, A.M., Gharabaghi, B., & McBean, E.A. (2016). Prediction of Timing of Watermain Failure Using Gene Expression Models. *Water Resources Management,* 30(5), 1635-1651. https://doi.org/10.1007/s11269-016-1241-x

Scheidegger, A., Leitão, J.P., & Scholten, L. (2015). Statistical failure models for water distribution pipes - A review from a unified perspective. *Water Research,* 83, 237-247. https://doi.org/10.1016/j.watres.2015.06.027

Shirzad, A., Tabesh, M., & Farmani, R. (2014). A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. *KSCE Journal of Civil Engineering,* 18(4), 941-948. https://doi.org/10.1007/s12205-014-0537-8

Tang, K., Parsons, D.J., & Jude, S. (2019). Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system. *Reliability Engineering and System Safety,* 186, 24-36. https://doi.org/10.1016/j.ress.2019.02.001

Tavakoli, R., Sharifara, A., & Najafi, M. (2020). Prediction of Pipe Failures in Wastewater Networks Using Random Forest Classification. *Pipelines*, December 2019, 90-102.

Wang, R., Dong, W., Wang, Y., Tang, K., & Yao, X. (2013). Pipe failure prediction: A data mining method. *Proceedings - International Conference on Data Engineering* (1208-1218). https://doi.org/10.1109/ICDE.2013.6544910

Wilson, D., Filion, Y., & Moore, I. (2017). State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal,* 14(2), 173-184. https://doi.org/10.1080/1573062X.2015.1080848

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., & Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees. *Structure and Infrastructure Engineering,* 14(10), 1402-1411. https://doi.org/10.1080/15732479.2018.1443145

Wu, J., & Baker, J.W. (2020). Statistical learning techniques for the estimation of lifeline network performance and retrofit selection. *Reliability Engineering and System Safety,* 200(March), 106921. https://doi.org/10.1016/j.ress.2020.106921