Application of regression tress for prediction of water conduits failure rate

Małgorzata Kutyłowska[1,*]

[1]Wrocław University of Science and Technology, Faculty of Environmental Engineering, Wybrzeże Wyspiańskiego 27, 50-370 Wroclaw, Poland

Abstract. This paper presents the results of predicting the failure rate of water distribution pipes and house connections in a selected city in Poland by means of regression trees. Several regression tree models were built as part of modelling. Optimal models were selected (separately for each of the water conduit types) via an analysis of the so-called costs. The regression tree structure comprised independent variables, i.e. predictors (length, diameter, year of construction and material). The failure rates of the two types of water conduits were the dependent variable. The optimal models were characterized by the lowest costs and a relatively simple tree architecture. Operational data from the years 2001–2012 were used to determine the experimental (real) values of the failure rate and to build regression tree models. The optimal models included eight divided nodes and nine end nodes. The ranking of the significance of the parameters showed that length was the predictor responsible for division on the successive tree levels. The obtained high consistency (0.99) of the real data with the predicted ones indicates that the regression tree method can be used to analyze and assess the failure rate of water conduits.

1 Introduction

Mathematical modelling has been applied in many spheres of life already for many years [1, 2]. Especially engineering problems are solved using mathematical tools [3, 4]. Environmental engineering is a field which has been developing very rapidly. Among the many branches of broadly understood environmental engineering one should distinguish the sphere of municipal management, i.e. water supply systems, sewage disposal systems and waste management systems. The three areas mentioned can be said to belong to the critical infrastructure since their role is absolutely vital for the security and quality of life of inhabitants [5, 6]. Many of the phenomena connected with operating hydraulics and the design of water distribution systems [7] or sewage disposal systems [8] have already been thoroughly explored and described in the literature on the subject. Therefore today increasing importance is attached to the proper management and operation of municipal systems. Hence it is necessary to use the available mathematical tools to forecast the dynamically changing parameters of water distribution network operation. Reliability indices, e.g. the failure rate, not only should be determined on the basis of operational data,

[*] Corresponding author: malgorzata.kutylowska@pwr.edu.pl

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

but also relatively easily and quickly implementable algorithms and models should be employed to predict such random variables. The failure rate has been usually determined on the basis of the available operational data [9, 10] or through model studies [11], statistical models [12, 13] and other less conventional modelling methods [14–16]. So far the regression tree methodology has not been widely used to assess the reliability of municipal system operation or to predict selected indicators. This is why the author decided to take up this subject. The main aim of this paper is to show the possibilities of applying the regression tree algorithm to the analysis of the frequency of failures of water conduits. This method has already found application in many other fields, e.g. in the assessment of the degree of damage to buildings [17], in the modelling of the rate of failure of pumping systems [18] and in broadly understood economy [19]. Thus it seemed worthwhile to find out whether it is suitable for assessing the frequency of failures of water supply networks.

Moreover, this paper is a complement to the research presented in another work by the author [20], in which the failure rate was predicted (for the same water distribution network) by means of artificial intelligence. Thus it will be possible to compare the effectiveness of the regression tree method with that of the artificial neural network algorithm.

2 Materials and methods

Regression trees and classification trees are used to predict respectively quantitative and qualitative variables. This method of data analysis and prediction began to be used in the 1960s, but in 1984 it was popularized by L. Breiman [21]. Generally speaking, a regression tree (RT) or a classification tree (CT) is a directed graph having a root and nodes (leaves), in which the conditions regarding variables are checked, and branches comprising decision rules. As a rule, the regression tree method is easier to implement and makes the analysis of the results easier than the classification tree method [21]. An analysis using a tree building algorithm consists in finding a set of logical division conditions, and relations between the predictors and the dependent variable, which leads to prediction results [21]. The advantage of using RTs and CTs is the relatively easy interpretation of the results and accurate predictions [22]. Moreover, regression tree models are resistant to outliers, which for various reasons may occur in the operational data acquired from water companies. When outliers appear, they are isolated in small nodes. If there are only a few such values, they can be neglected [21].

The failure rate (λ, fail./(km·a)) of water supply pipes was predicted using the regression tree method. Since the failure rate of the distribution pipes (λ_r) and that of the house connections (λ_p) were to be predicted separately it was necessary to build two different tree models. Rates λ were the dependent variables while the predictors (independent variables) were: length, diameter, year of construction and material, separately for the distribution pipes and the house connections.

Operational data for the years 2001–2012, obtained from a water company in one of the medium-sized cities in Poland, were used to determine the actual failure rate λ and to predict it by means of regression trees. When building a regression tree model, one should bear in mind that too complex trees are difficult to interpret. Therefore, as in any other approach so in the case of the RT method, the aim is to create as simple models as possible, but fully rendering the interrelations between the dependent variables and the predictors. The values of the experimental dependent variables and the predictors are presented in table 1.

Table 1. Dependent variables and predictors.

Length, km	**Diameter, mm**	**Year of construction**	**Material**	**λ, fail./(km·year)**
House connections				
23.4–50.2	20–100	1961–2012	cast iron, steel, galvanized steel, PE, PVC	0.23–1.59
Distribution pipes				
57.3–88.7	80–200	1961–2006	cast iron, steel, PE, PVC	0.10–0.57

The optimal regression tree model was selected on the basis of the resubstitution of costs, where the expected squared error is calculated from the relation [21]:

$$R(d) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - d(x_i)\right)^2 , \tag{1}$$

where training sample Z consists of points (x_i, y_i), for i = 1, 2,..., N. The calculations are performed for the same data set which was used to build model d [21]. In this case training sample Z consists of such independent variables (x) as length, diameter, material and year of construction of water pipe as well as dependent variable (y) – failure rate (see table 1). A regression tree is created through iterative divisions in the nodes so as to minimize the cost [21]. The notion of cost (in the RT method [22]) is a generalization of the idea that the model with the smallest error is characterized by the best prediction. The measure of cost is a ratio of the incorrectly defined cases to all the cases. Thus the optimal model should be characterized by the lowest cost. An important element of the analysis and the choice of tree size is *V*-fold cross-validation, consisting in dividing the data into *V* randomly selected disjoint parts. Using the remaining (*V-1*) parts of the data as training cases, the dependent variable is predicted and the prediction error is calculated. The tree of a given quantity is calculated *V* times. Each time cases with one subsample excluded (to serve as a test sample) are used in the calculations. Thus each of the subsamples is used *V-1* times in the training sample and only once as the test sample. The cross-validation cost (CV cost) is calculated as the average cost for *V* test samples. This average is a CV cost estimate [22]. In this study 10-fold cross-validation was used. A given tree must be created many times and a large data set is needed to perform the divisions mentioned above. The tree structure (the number of branches and nodes) depends on the number of divisions responsible for the best prediction. Divisions are performed until the nodes are homogenous or comprise a specified number of cases. The computations presented in this study were carried out using the Statistica 12.0 software.

3 Results and discussion

Ten and twelve regression tree models were built to predict respectively failure rates λ_r and λ_p. The methodology described above was used to select optimal regression tree models. Their structures are shown in figure 1. There were 8 divided nodes and 9 end nodes in each of the models for predicting the failure rate of respectively distribution pipes and house connections. Under close examination the models differ considerably in their architecture. Differences are visible already on the third level of division, i.e. in nodes (leaves) number 4, 5, 10 and 11. One can say that the tree model for distribution pipes is more complicated since most (6 out of 9) of the end nodes are on the last level of division, which may indicate that until the very end of division the decision rules were not explicit. The architectures of the regression trees (fig. 1) do not seem to be too complex, considering the number and type of independent variables which make up the model. It should be mentioned that regression trees can be connected into assemblies of trees to form a random forest. As a rule, an assembly of trees gives better prediction results than a single, even most complicated, tree [22]. In the case of a random forest, it would be necessary to include many other variables in the vector of predictors so that the complicated model architecture reflected the relations between the predicted variable and the independent variables. This is naturally limited by the acquirability of many operational variables. It seems that information on the pressure prevailing in the water supply network and other data on the pipelines (e.g. the number and kind of failures) and even data seemingly not connected with the failure rate problem, such as water production and demand or losses, should be added to the vector of predictors.

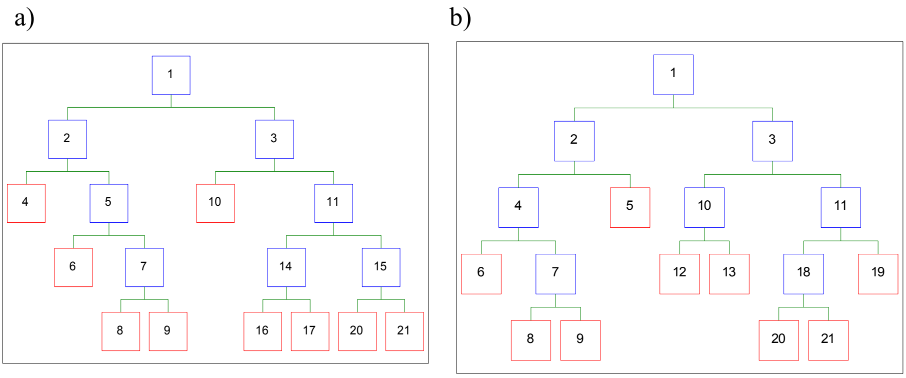

Fig. 1. Optimal structure of regression tree, a) distribution pipes, b) house connections.

Cost sequence diagrams after cross-validation for the RT models are shown in figure 2. Models no. 3 and 4, for which the costs amounted to 0.000333 and 0.001086, were selected as optimal for the prediction of rate λ_r and λ_p, respectively.

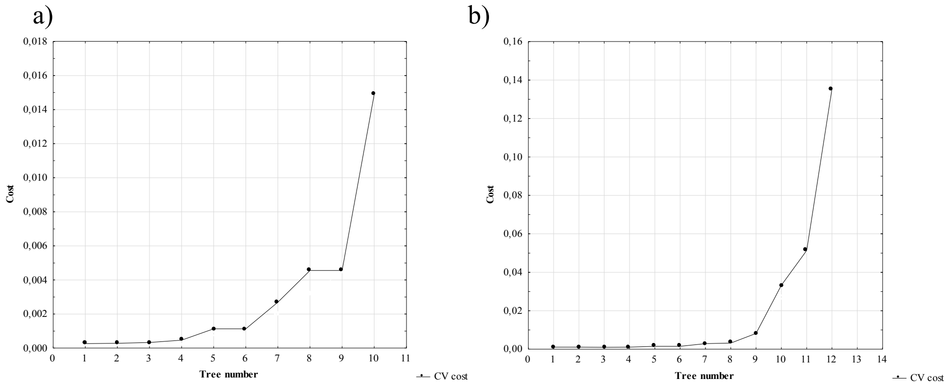

Fig. 2. Cross-validation cost (CV), a) distribution pipes, b) house connections.

It was found that the costs would increase as model complexity decreased, e.g. models 10 and 12 had only one end node. On the other hand, when selecting an optimal model one should also take into account the simplicity of the regression tree architecture. It is not a great achievement when one selects a model characterized by very low resubstitution cost values, but whose structure is very complicated (e.g. there are 10–20 end and divided nodes).

For the division into branches and levels it is essential to determine the so-called importance, i.e. rank the predictors according to their significance on the scale of 0–1. This approach is helpful in identifying the variables having a significant prediction power with regard to dependent variables [21, 22]. When predicting the failure rate of house connections (λ_p), the independent variables were ordered as follows: conduit length, material, year of construction and diameter. The importance of the above predictors amounted to respectively: 1.00, 0.07, 0.03 and 0.01. In the case of the regression tree used to model rate λ_r, the significance ranking was a little bit different: pipeline length (1.00), year of construction (0.08), diameter (0.04) and material (0.02). It is apparent that the conduit length predictor dominated in the optimal models, considering that its value was by two orders of magnitude higher than the values of the other independent variables. This fact is quite inexplicable since it is rather such independent variables as conduit material, diameter and laying year should be the predictor having a greater influence on the prediction of the dependent variable (the failure rate). The research so far has shown that not pipeline length, but the other variables are relatively important in the assessment of the failure rate of water conduits [23, 24]. However, an analysis of figure 3, illustrating the real (experimental) and predicted (by the RT model) failure rates of respectively the house connections and the distribution pipes, shows that despite the quite surprising dominance of one dependent variable (length) the modelling results are simply ideal. The regression tree method considerably differs from other regression algorithms, such as support vector machines or K-nearest neighbours, which may affect the quality of modelling and the choice of a particular independent variable as the dominant parameter in the prediction of the failure rate of water conduits.

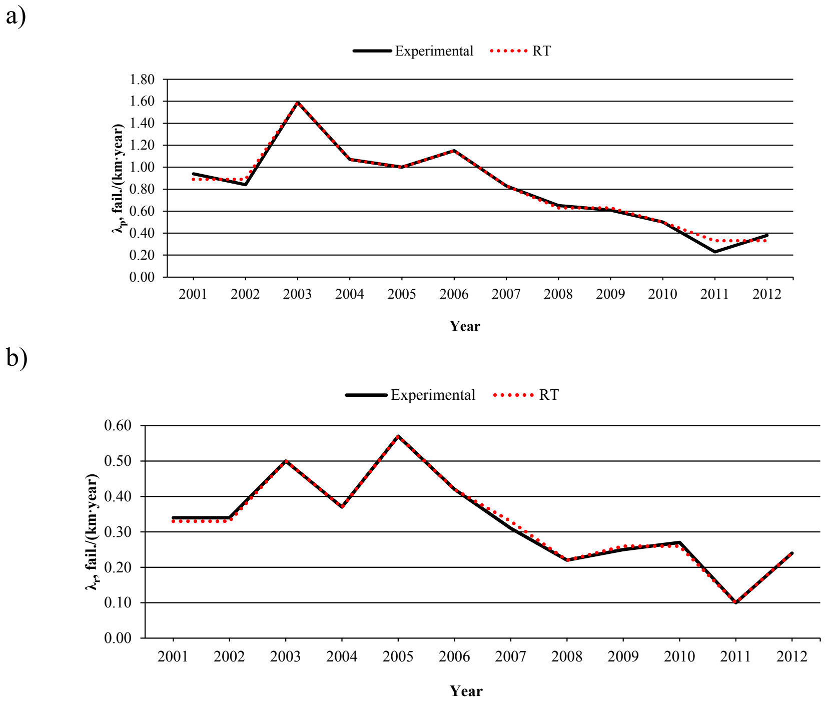

Fig. 3. Experimental and predicted failure rates, a) house connections, b) distribution pipes.

Almost ideal convergence between the real failure rate and the predicted one was observed for the house connections (fig. 3a). Coefficients R and R^2 amounted to respectively 0.994 and 0.988. Only in 2011 the predicted failure rate was slightly higher than the real one. The Spearman rank correlation, which is a nonparametric measure of statistical interdependence between variables, is also high, amounting to 0.995. When the RT modelling results are compared with the results of predicting rate λ_p by means of artificial neural networks (ANN) [20], it becomes apparent that the latter method generates greater divergences between experimental and predicted data. However, one should note that the ANN modelling methodology is somewhat different the RT methodology. In the case of the RT algorithm, a whole data set was used to build the model. Whereas ANN models are created on the basis of a data set which is divided into two subsets: a training subset and a testing subset. The prediction results for the testing set are characterized by larger real-predicted value errors due to the fact that the ANN model previously "saw" only the training set data. Therefore it seems necessary to broaden the present research by carrying out the validation of the RT model on the basis of the operational data from the next years. In this way one can conclusively determine whether the regression method is more advantageous as regards the correct prediction of the failure rate of water supply pipelines.

The prediction of the rate of failure of the distribution pipes (fig. 3b) is very good. The relative prediction error amounted to about max. 6.5%, which is a satisfactory result. The correlation between the experimental results and the ones predicted by the RT method was $R^2 = 0.996$. The Spearman rank correlation is at a similar level as in the analysis of the failure rate of the house connections, amounting to 0.993. The rank correlation values indicate that the variables (the experimental and the predicted failure rates) form an increasing function. The analyses show that the RT algorithm is quite a universal approximator. The other regression methods, e.g. artificial neural networks [20], are not always able to make a correct prediction of data slightly divergent from the other values in the set (e.g. the values of rate λ_r in 2011). As mentioned above, the resistance of the RT method to outliers is undoubtedly its advantage over the artificial intelligence algorithm.

However, when analyzing the suitability of regression methods for predicting failure rates one should take into account not only the agreement between the predicted data and the real ones, but also the type of pipeline and its influence on the reliable operation of the whole water distribution network. Because of the higher costs of repair and the social costs resulting from a drop in network pressure, interruptions in water supply and other failure events, failures of a trunk water main or a distribution pipe have more significant consequences than a failure or 10–20 house connections at the same time. Since the prediction results are promising, it seems that in the future the regression tree method should be considered as an alternative to other regression methods for predicting and analyzing the failure rate of water supply pipelines.

Failure rate prediction using mathematical modelling, e.g. regression trees or other regression methods as artificial intelligence or support vector machines, could be useful for water companies. The modelling results could be used by water utilities to e.g. plan the modernization schedule or to propose the proper renovation method for specific water pipe section. In such case including the street name and exact localization of occurred damage seems to be reasonable and useful especially for exploiters who are interested a lot of in practical and technical aspects. We see that at this stage there many additional facts and data which should be taken into consideration during forecasting of failure frequency. Moreover, this kind of failure rate assessment (by means of RT) allows us to distinguish and choose the most important variables (see significance ranking described above). This knowledge is also very useful for water companies, because in relatively simple way we point out what kind of data and with what accuraccy should be registered.

4 Conclusions

The results of predicting the rate of failures of the distribution pipes and house connections in one of the Polish cities by means of the regression tree method have been presented. The subject seems to be important for the correct and quick estimation of the reliability level of municipal systems. The created RT models can be useful in cases when it is necessary to determine the failure frequency in order to make a decision on planned repairs of water supply pipelines. The modelling methodology presented here is something of a novelty in comparison with the approach used so far to predict the failure rate of water conduits. A survey of literature on this subject revealed that the regression tree method is not widely used to assess the level of operational reliability of water supply networks, which induced the author to take up this subject. The main conclusions emerging from the modelling are as follows:

- ideal convergence (at the level of 0.99) between the experimental data and the values predicted by the regression tree models was obtained for both the distribution pipes and the house connection;

- the optimal RT models, selected on the basis of a minimal cost analysis, comprised 8 divided nodes and 9 end nodes; despite the fact that the number of nodes was the same in both the model for predicting the failure rate of house connections and the one for predicting the failure rate of the distribution pipes, the architecture of the two models was considerably different: different nodes (leaves) would undergo division and the number of nodes on the particular tree levels was different; the tree model for the failure rate of house connections seems to be less complicated since its final nodes are uniformly distributed between the last and the last but one level of the tree, which may indicate that the final divisions were performed uniformly and the decision rules were quite explicit;

- the lowest costs were recorded for the models the structures of which were quite complicated (10–20 divided and end nodes); the criterion for selecting the optimal model cannot be based on costs only since the aim is also to minimize model complexity; as regards regression trees, it is better to replace the very complicated structure of a single tree with a random forest; this seems to be sensible for predicting random variables, which reliability indices undoubtedly are;

- the analysis of the ranking of significance showed that length is an independent variable having the greatest influence on the tree structure; the importance of this parameter amounted to 1.00 and that of the other parameters was lower by as many as two orders of magnitude; it was precisely the length of pipelines which was responsible for divisions in the nodes on the successive levels of the tree;

- the comparison of the results of modelling by means of respectively regression trees and artificial neural networks [20] indicates that the latter method is less resistant to outliers, whereby the convergence between the predicted values and the real ones is lower.

The work was realized within the allocation No. 0401/0006/17 awarded for Faculty of Environmental Engineering Wroclaw University of Science and Technology by Ministry of Science and Higher Education in years 2017–2018.

References

1. Q. Meng, C. J. Cieszewski, M. Madden, Gisci. Remote Sens. **44**, 2 (2007)
2. I. Kasprzyk, A. Grinn-Gofroń, A. Strzelczak, T. Wolski, Sci. Total Environ. **409**, (2011)
3. F. Ünes, M. Demirci, Ö. Kisi, Period. Polytech-Civ. **59**, 3 (2015)

4. B. Tchórzewska-Cieślak, K. Pietrucha-Urbanik, Eksploat. Niezawodn. **18**, 2 (2016)
5. B. Tchórzewska-Cieślak, Global Nest J. **16**, 4 (2014)
6. Y. Omar, A. Parker, J. Smith, S. Pollard, Sci. Total Environ. **576**, (2017)
7. E. Avila-Melgar, M. Cruz-Chávez, B. Martinez-Bahena, Wat. Sci. Technol. **16**, 6 (2016)
8. B. Kaźmierczak, M. Wdowikowski, Period. Polytech-Civ. **60**, 2 (2016)
9. M. Kutyłowska, M. Orłowska-Szostak, Water Practice and Technology **11**, 1 (2016)
10. M. Bakker, E.A. Trietsch, J.H.G. Vreeburg, L.C. Rietveld, Wat. Sci. Technol. **14**, 6 (2014)
11. M. Iwanek, D. Kowalski, M. Kwietniewski, Ochr. Sr. **37**, 4 (2015)
12. H. Osman, K. Bainbridge, J. Perform. Constr. Fac. **25**, 3 (2011)
13. A. Scheidegger, J.P. Leitao, L. Scholten, Water Res. **83**, (2015)
14. M. Aydogdu, M. Firat, Water Resour. Manag. **29**, 5 (2015)
15. K.N. Fleming, B.O.Y. Lydell, Reliab. Eng. Syst. Safe. **86**, 3 (2004)
16. G. Kabir, G. Demissie, R. Sadiq, S. Tesfamariam, Knowledge-Based Systems **85**, (2004)
17. A. Malinowska, Nat. Hazards **73**, 2 (2014)
18. M. Bevilacqua, M. Braglia, R. Montanari, Reliab. Eng. Syst. Safe. **79**, 1 (2003)
19. A. I. Irimia-Dieguez, A. Blanco-Oliver, M. J. Vazquez-Cueto, Procedia Economics and Finance **23**, (2015)
20. M. Kutyłowska, Eng. Fail. Anal. **47**, (2015)
21. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression tress (Chapman & Hall/CRC, Boca Raton, USA, 1984)
22. Statistica 12.0, Electronic Manual
23. K. Shahata, T. Zayed, Struct. Infrastruct. E. **8**, 11 (2012)
24. G. Pelletier, A. Mailhot, J.P. Villeneuve, J. Water Res. Pl-ASCE. **129**, 2 (2003)