DOI: 10.2428/ecea.2017.24(2)17 ECOL CHEM ENG A. 2017;24(2):193-205

Małgorzata KUTYŁOWSKA[1]

REGRESSION METHODS FOR PREDICTING RATE AND TYPE OF FAILURES OF WATER CONDUITS

METODY REGRESYJNE DO PRZEWIDYWANIA POZIOMU AWARYJNOŚCI I RODZAJU USZKODZEŃ PRZEWODÓW WODOCIĄGOWYCH

Abstract: This paper demonstrates that regression trees (RT) and classification trees (CT) can be applied to predict the rate and type of failures of water conduits. An analysis by means of a tree building algorithm consists in finding a set of logical division conditions and determining correlations between the predictors (independent variables) and the dependent variable, in consequence of which prediction results are obtained. The failure rate of distribution pipes (DP) and house connections (HC) was predicted on the basis of operational data for the years 2008–2014 for one water supply zone of a medium-sized Polish city. The independent variables were: the length of a particular type of conduits and the number of DP and HC failures recorded in a particular year. Separate regression tree models were created for modelling the failure rate of respectively DP and HC. In the case of the classification problem, one model was built for jointly DP and HC failures. In this model the qualitative dependent variable was type of failure while the predictors were material and conduit diameter and type. The results indicate that the RT method can be used to evaluate the failure rate of water conduits. Whereas the classification of failure types was not fully satisfactory, which means that further research in this area is needed. The calculations were performed using Statistica 13.1.

Keywords: regression methods, water supply network, kind of damage of water conduits

1. Introduction

Today mathematical modelling is an indispensable tool for solving many complicated engineering problems. Knowledge, experience and intuition are used to build mathematical models of random experiments. But prior to modelling one should analyse the experimental (operational data) to be used to create a model. Since it is a highly subjective part (combining to some extent science and art) of the cognitive work, the analysis of the data can lead to quite different conclusions. Bearing this in mind, the operational data obtained from water companies for the purpose of building of models

[1] Faculty of Environmental Engineering, Wroclaw University of Science and Technology, Wybrzeże St. Wyspiańskiego 27, 50-370 Wrocław, Poland, phone: +48 71 320 40 84, email: malgorzata.kutylowska@pwr.edu.pl

should be properly handled and subjected to a qualitative and quantitative analysis. This paper presents the application of a selected modelling method – the regression method (classification and regression trees) – to the prediction of the failure rate of water conduits and the type of damage to the latter.

1.1. Regression and classification tress

Classification and regression trees have been used for predicting respectively qualitative and quantitative variables. This method of analysing and predicting data began to be used in the 1960s, but it was as late as 1984 when it was popularized by Breiman [1]. Generally speaking, a regression tree (RT) or a classification tree (CT) is a directed graph comprising a root and nodes (leaves), in which conditions applying to the variables are checked, and branches containing decision rules. As a rule, it is easier, in comparison with the classification method, to implement the regression tree method and analyse its results [1]. An analysis by means of a tree building algorithm consists in finding a set of logical division conditions and determining correlations between the predictors and the dependent variable, in consequence of which prediction results are obtained [1]. The advantage of using trees is that the results of prediction are good and they can be relatively easily interpreted [2]. Moreover, regression tree models are resistant to outliers, which often and for different reasons crop up in the operational data obtained from water companies. If outliers appear, they are isolated in small nodes. If there are not many of them, they can be omitted [1]. The structure of a tree (the number of branches and nodes), ensuring the best prediction, depends on the number of divisions. Divisions are made until the nodes are uniform or comprise the specified number of cases. The optimal regression tree model is selected on the basis of the resubstitution cost, where the square error is calculated from the relation [1]:

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

in which the training samples consists of points (x_i, y_i) at $i = 1, 2, ..., N$. The calculations are performed for the same data set which was used to build model d [2]. A regression tree is created through iterative divisions in the nodes to minimize the cost [1]. The notion of "cost" in the CT and RT methods [2] is a generalized idea that a model with the smallest error yields the best predictions. A measure of the cost is a ratio of the incorrectly defined cases to all the cases. Thus an optimal model should be characterized by the lowest cost. Moreover, V-fold cross validation should be carried out in order to select a proper tree size. The cost of the cross validation is calculated as the average cost of the V test samples. This average is an estimate of the cross validation cost [2]. It is required that a given tree be created many times and that the data set be adequate for making the above mentioned divisions. For the division into successive branches and levels it is essential to determine the so-called importance, i.e. rank the significance of the predictors using the scale of 0–1. This approach is helpful in identifying independent variables with significant prediction power towards the dependent variables [1, 2].

Today regression and classification trees are used in many fields of broadly understood environmental engineering. For example, the failure frequency of the pumping systems used in the refining industry was modelled by means of RTs and CTs [3]. This way of modelling was found to contribute to greater reliability of the whole pumping system, which is highly important since the refining industry (similarly as water distribution systems) belongs to the critical infrastructure. Regression trees can be connected into complexes to form a distributed random forest (DRF). The latter was used to predict the pollution of underground water with nitrates [4]. Sun et al. used DRF to model the amount of solar radiation depending on the degree of air pollution [5], noting the ability of DRF to indicate the significance ranking of the independent variables used to build the model. Wang et al. proposed to use DRF to estimate the risk of flooding [6]. Malinowska [7] tested the suitability of classification and regression tree modelling for estimating the risk of damage to buildings in mining damage areas in Upper Silesia. A survey of the available world literature on the subject shows that also financial problems [8] and the probability of an accident occurring on a motorway [9] can be modelled using the RT and CT methods.

1.2. Water-pipe network

At the present stage of development, when the methods of designing water supply systems have been sufficiently well verified and when hydraulics (generally speaking, fluid mechanics) is a well-known and widely applied discipline, it seems that in the case of water distribution systems the emphasis should be placed on their upgrading and proper operation (taking into consideration also the dependence between water quality and the pipe material) [10] and on research on the operational reliability and proper management of water supply [11] in order to extend the life and failure-free operation of the underground facilities. The operation of each element of a water supply system requires an individual approach to the description and modelling of the phenomena taking place in a given facility, taking into account the latter's function. For example, the approach to the modelling of pressure variation and flows in a water supply network [12] should be different than the one (e.g. employing artificial intelligence) used to select water conduits for rehabilitation [13]. Problems relating to the operational reliability and failure frequency of water supply networks and water losses have been investigated by many research teams in Poland [14–17], which has contributed to the development of this science, including its modelling aspect, not only in Poland, but also abroad [18, 19]. This paper is an attempt to supplement the above research with the modelling of reliability indicators (using the water conduit failure rate as an example) by means of regression methods (machine learning methods to which regression and classification trees belong). As a modelling tool RTs and CTs are widely used in many fields, but practically until now no application of this methodology to the analysis of the failure frequency level and the prediction of the failure rate of water conduits has been reported in the world and domestic literature. This induced the author to undertake this subject. Regression trees are used here to model the failure rate (a quantitative dependent variable) while classification trees are employed to predict the type of failure (a qualitative dependent variable).

2. Materials and methods

The failure rate, λ [fail./(km·year)] of water conduits was predicted using the regression tree method. It was predicted separately for distribution pipes (DP) – λ_r and house connections (HC) – λ_p, which meant that two different tree models had to be built. Rates λ were the dependent variables while the predictors (independent variables) were: the length of the conduit of a particular type (L_p and L_r) the number of failures (N_p and N_r) of respectively HC and DP, recorded in a given year. Conduit length and number of failures were selected as the basic variables since such data are definitely recorded by water companies and so are easily available. As part of this research also the suitability of this basic information (lacking details on the pipe material and diameter) for predicting the failure rate by means of regression trees was verified.

Moreover, classification trees were used to predict the type of failure. In this case, one common model was created for both distribution pipes and house connections. The vector of qualitative independent variables comprised: the material – *M* (cast iron, steel, PVC and PE) and the type of conduit – *T* (DP and HC). The quantitative predictor was the diameter – *D*. The predicted qualitative dependent variable was the type of failure – *R* (corrosion, hole, longitudinal fracture and lateral fracture). When building CT models three kinds of goodness of fit: the Gini coefficient, the Chi-square statistic and the G-square statistic were used as well as 10-fold cross validation was applied.

Operational data for the years 2008–2014 obtained from a water company in one of the Polish medium-sized cities were used to determine the real failure rate λ and to predict the failure rate and type of damage by means of the regression and classification tree method. The whole water distribution system was divided into 55 supply zones. This particular analysis focused on one selected water supply zone in which the pipeline pressure amounted to 0.4 MPa. The water pipe network in this zone supplies water to the inhabitants of one borough within the city. The overall length of the distribution pipes and the house connections was constant over the considered period, amounting to 31.7 km. The length of the distribution pipelines, made of grey cast iron (48.6%, 8.5 km), PVC (38.9%, 6.8 km) and PE (12.5%, 2.2 km), amounted to 17.5 km. In total there were 599 house connections with an overall length of 14.2 km. Water is supplied to the considered zone (with the total area of 41 km^2 and a population of about 10 000) from the so-called auxiliary drinking water intake, i.e. a 100 m deep well from which 1920 m^3 of water per 24 h are drawn from Upper Jurassic deposits.

The values of the dependent variables and predictors for the years 2008–2014 are given in Tables 1 and 2.

Table 1

Predictors and dependent variables – RT method

L_r [km]	L_p [km]	N_r [fail.	N_p [fail.]	λ_r [fail./(km·year)]	λ_p [fail./(km·year)]
17.5	14.2	2–5	3–10	0.11–0.29	0.21–0.70

Table 2

Predictors and dependent variables – CT method

M	*T*	*D* [mm]	*R*
Cast iron, steel, PVC, PE	Distribution pipe, house connection	25–250	Corrosion, longitudinal fracture, lateral fracture, hole

3. Results and discussion

The calculations were performed using Statistica 13.1. Several regression tree models for predicting failure rates λ_r and λ_p and failure type *R* were built. Optimal RT and CT models, whose structures are shown in Figs 1 and 2, were selected. In the Fig. 1 the average value (Av) and variance (Var) of independent variable were presented. The model most suitable for predicting the failure rate of the distribution pipes and the house connections was selected taking into account: the lowest resubstitution cost, the small degree of model complexity and the quality of prediction, i.e. the convergence of the real (experimental) dependent variable with the values obtained from modelling.

The architectures of the two models (Fig. 1) seem to be the same. Actually, the number of divided nodes (1 node) and that of end nodes (2 nodes) in the two models is the same, but since the values of the independent variables are different, the average and variance values considerably differ between the models (Fig. 1a and 1b). Moreover, a different value is responsible for the division of the divided node into two end nodes.

In the case of the classification problem, the selection of the model most suitable for predicting the type of failure was based on not only the comparison of the costs, but also on the number of incorrectly classified cases and the degree of model complexity. The model using the Gini coefficient was selected. The architecture of a CT model (Fig. 2) is a bit more complex than that of an RT model since it includes 2 divided nodes and 3 end nodes.

Independent variable *M*, i.e. the conduit material, was responsible for the division into the successive tree levels. On the first level, it was steel and all the other kinds of material while on the next level it was PVC. One should note that corrosion and fracture were the dependent variables which dominated in the particular nodes of the CT model. This means that the failure type “hole” was not of much importance for determining the quality of the model. This could be due to the fact that this dependent variable was the least numerous one.

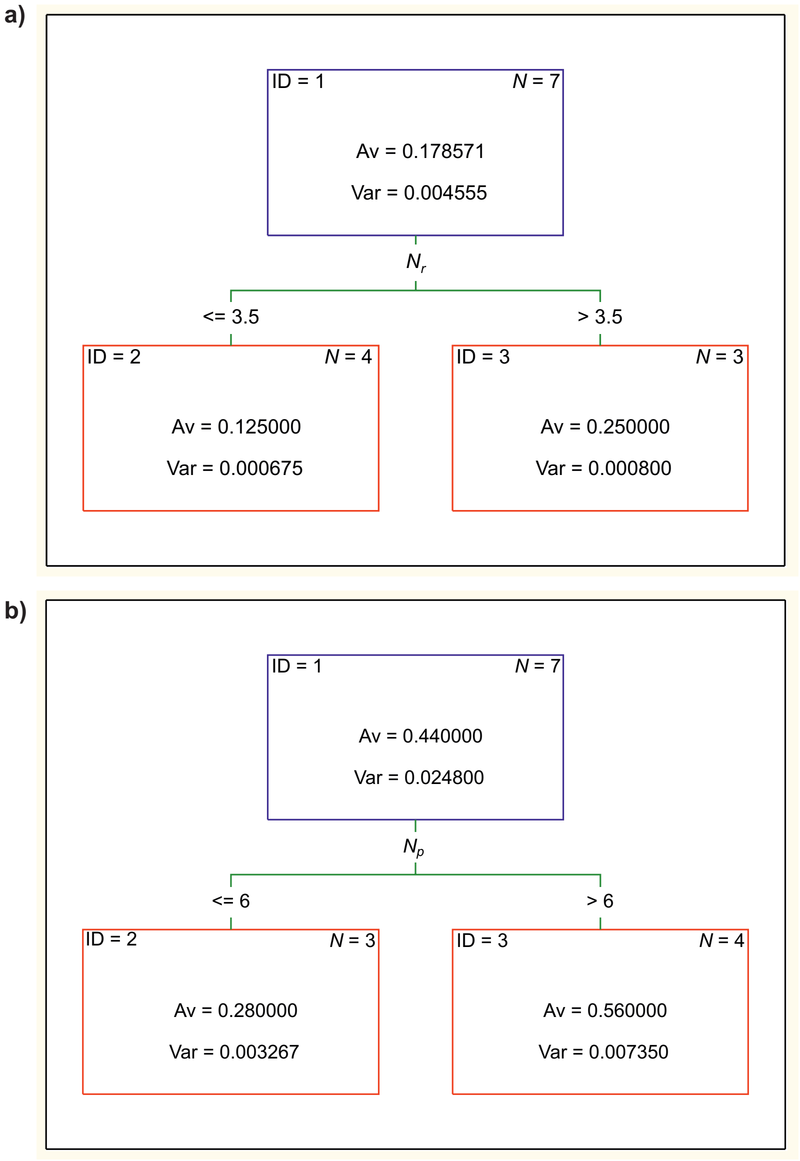

Fig. 1. Optimal regression tree structure: a) distribution pipes, b) house connections

Resubstitution costs were used to evaluate model quality, but in the case of the CT model, also the cost of cross validation was taken into account. The cost amounted to 0.00073 for the RT model describing the failure rate of the distribution pipes and to 0.0056 (a value by one order of magnitude higher) for the model describing the failure rate of the house connections. For the other models (other than the optimal model) the costs would linearly increase with the number of divided nodes and end nodes (i.e. with

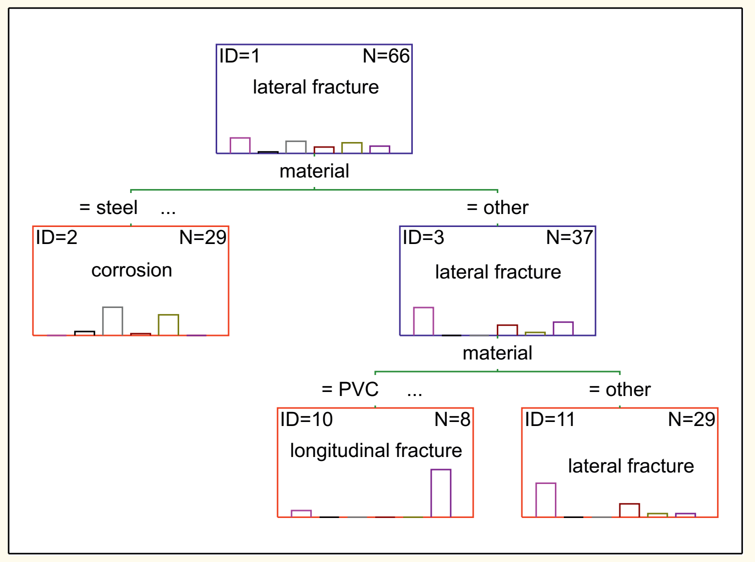

Fig. 2. Optimal structure of classification tree

increasing model complexity) for both DP and HC. A comparison of the resubstitution costs for the selected RT models for predicting the failure rate indicates that despite the same regression tree architecture, the models considerably differ in their modelling quality (expressed by the resubstitution cost value). As regards the selected optimal model for classifying the type of failure, the resubstitution cost amounted to 0.39394 and the cross validation cost to 0.40909. For the other CT models the values were higher, except for one tree model which was characterized by the resubstitution cost of 0.33333 and the cross validation cost of 0.39394. These values are a little lower than in the case of the selected optimal CT model, but the tree structure was highly complex as it had 5 divided nodes and 6 end nodes. As already mentioned, when selecting a model one should consider not only the quality of predicted data and the costs incurred to build the model, but also the model architecture simplicity which makes for the easier interpretation of the results and a greater generalization capability. In the case of more complex models there is a danger of a too close fit between the real data and the predicted ones, whereby the model loses its ability to adapt to changing conditions which may occur when an independent variables vector other than the one used to build the model is included in the latter.

Also the independent variables were compared with regard to significance ranking. For the two optimal RT models the number of failures in a given year (N_p and N_r) was the independent variable with the highest importance of 1.00. This is shown in Fig. 1 where the number of failures is the variable responsible for the division into the successive regression tree levels. It emerges from the analysis that the length of conduits actually does not affect the building and quality of RT models. This can be due to the fact that the vector of predictors was not numerous, i.e. it consisted of only two variables. Quite a different conclusion emerged from an analysis of the results of modelling the failure rate in another Polish city by means of regression trees [20]. In this case, the length of pipelines was the dominant variable. However, the difference can stem from another independent variable vector used to build the model. Besides conduit length, such variables as: diameter, year of construction and material were the predictors in the above work [20]. Nevertheless, in the present paper such basic independent variables as *N* and *L* were used on purpose (which should be emphasized) in order to find out if regression trees could be used to model the failure rate of water conduits even when little information on a considered water distribution system is available. The significance of the particular independent variables for the CT model is illustrated in Fig. 3. The most important predictor was material. This is also true for engineering practice since in many cases the type of failure is closely connected with the material from which the conduit is made. For example, material corrosion will never occur in the case of a pipeline made of plastic. Diameter and type of conduit had a similar (relatively high) significance in the considered classification problem.

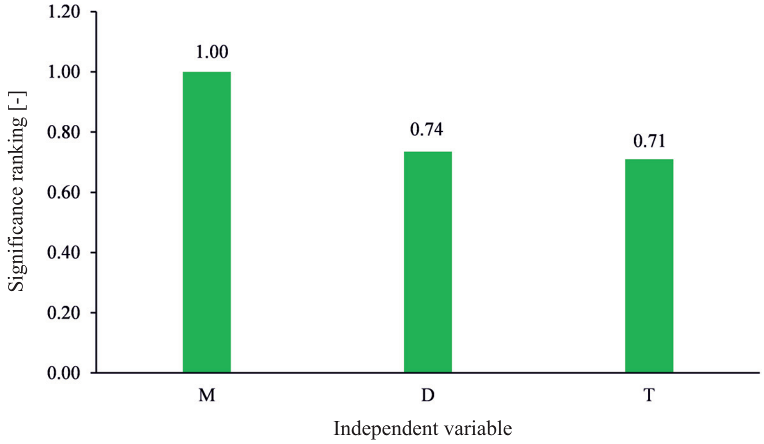

Fig. 3. Significance ranking of the predictors in CT model

An analysis of the rate-of-failure prediction results for the distribution pipes and the house connections (Figs 4 and 5) shows that the use of only two variables in the predictor vector did not affect the quality of modelling. Even though the results are not so perfectly convergent as in [20], they are satisfactory from the engineering point of view. The maximum absolute error of failure rate modelling amounted to 0.04 fail./(km·year) and 0.14 fail./(km · year) for respectively DP and HC. Figure 4 shows that for the five of the seven analysed years the modelled failure rates are slightly higher than the real ones. In the case of distribution pipes such a small overestimation does not raise doubts as to the quality of modelling since a predicted failure rate higher than the real one can only induce the network operator to decide to renovate or replace selected network segments.

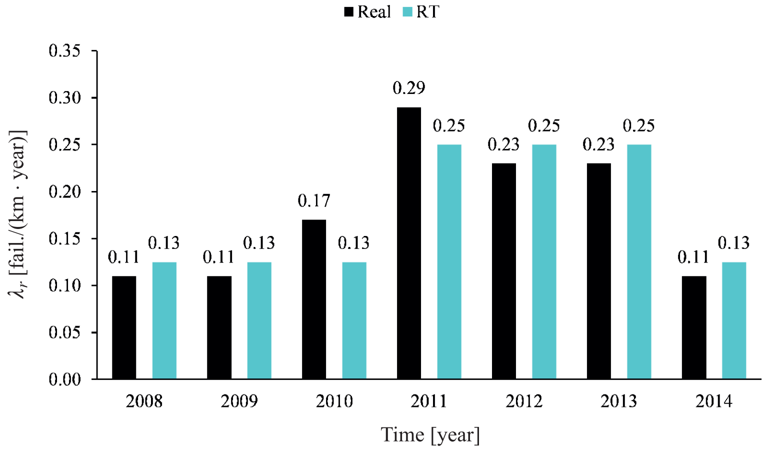

Fig. 4. Real and predicted values of failure rate (λ_r) of distribution pipes

One should bear in mind that the rank of distribution pipes is higher than that of house connections, which means that the small underestimation of the failure rate for the years 2009 and 2012 (Fig. 5) cannot provide ground for the statement that RT models are not a good tool for predicting the failure rate of house connections. It should be noted, however, that the results presented in this paper are for the prediction of failure rates λ_r and λ_p on the basis of the training sample, i.e. the data sample used to build the model. Therefore it seems reasonable to continue research on methods of building regression tree models and applying them so that at a later stage their quality could be verified using independent variables not included in the analysis before.

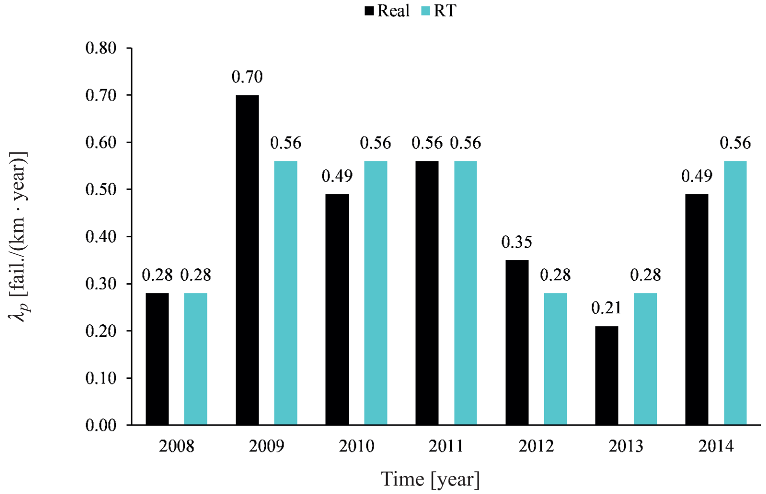

Fig. 5. Real and predicted values of failure rate (λ_p) of house connections

Table 3 shows the recorded types of failure and the ones predicted by means of the selected optimal CT model. Over the 7 years of operation the total of 44 house connection failures (of all types) occurred. 15 holes, 8 longitudinal fractures and 6 lateral fractures were registered and in 15 cases a pipe was corroded. 22 failures: 13 lateral fractures and 9 longitudinal fractures affected the distribution pipes. Owing to the fact that a single classification tree model was built on the basis of data for DP and HC jointly (as opposed to failure rate modelling for which two different models were created), Table 3 shows all the failures for the years 2008–2014.

Table 3

Registered and predicted types of damages

R – observed	R – predicted	R – observed	R – predicted
lateral fracture	lateral fracture	corrosion	corrosion
lateral fracture	lateral fracture	longitudinal fracture	lateral fracture
lateral fracture	lateral fracture	corrosion	corrosion
lateral fracture	lateral fracture	longitudinal fracture	longitudinal fracture
lateral fracture	lateral fracture	corrosion	corrosion
hole	corrosion	corrosion	corrosion
hole	corrosion	lateral fracture	lateral fracture
hole	corrosion	longitudinal fracture	pęknięcie poprzeczne
hole	corrosion	lateral fracture	lateral fracture
corrosion	corrosion	corrosion	corrosion
corrosion	corrosion	hole	corrosion
hole	corrosion	longitudinal fracture	longitudinal fracture
longitudinal fracture	corrosion	longitudinal fracture	longitudinal fracture
hole	corrosion	longitudinal fracture	longitudinal fracture
hole	corrosion	longitudinal fracture	longitudinal fracture
lateral fracture	lateral fracture	hole	corrosion
lateral fracture	lateral fracture	longitudinal fracture	longitudinal fracture
lateral fracture	lateral fracture	longitudinal fracture	lateral fracture
hole	corrosion	lateral fracture	longitudinal fracture
hole	lateral fracture	corrosion	corrosion
longitudinal fracture	lateral fracture	corrosion	corrosion
corrosion	corrosion	corrosion	corrosion
hole	corrosion	lateral fracture	lateral fracture
hole	corrosion	lateral fracture	lateral fracture
hole	corrosion	corrosion	corrosion
corrosion	corrosion	longitudinal fracture	lateral fracture
hole	lateral fracture	lateral fracture	lateral fracture
corrosion	corrosion	lateral fracture	lateral fracture
longitudinal fracture	lateral fracture	lateral fracture	lateral fracture
longitudinal fracture	longitudinal fracture	lateral fracture	lateral fracture
longitudinal fracture	lateral fracture	longitudinal fracture	lateral fracture
lateral fracture	lateral fracture	corrosion	corrosion
longitudinal fracture	lateral fracture	lateral fracture	lateral fracture

The failures incorrectly classified by the CT model are marked red in Table 3. Altogether there were 26 such cases from the total of 66 recorded failures in the considered water supply network, which amounts to over 39%. This is not a perfect result. Therefore it seems that further research on the application of classification trees to the modelling such a qualitative parameter as the type of failure of distribution pipes and house connections is necessary. In the case of the other statistics, i.e. the Chi-square statistic and the G-square statistic, the number of incorrectly classified cases was almost identical, amounting to respectively 26 and 24. This means that the type of fit statistic has no major effect on the quality of the model and its classification capability.

It is worth noting that in half of the incorrectly classified cases the model confused damage *hole* with pipeline material corrosion. This can be due to the fact that an independent variable having the same value, i.e. a house connection diameter of 32 mm or 40 mm, was associated with the two types of inoperability. The worse results of modelling (quite many incorrectly classified cases) by the CT method in comparison with the RT method, where the failure rate was predicted reasonably correctly (with an error admissible from the engineering point of view), can be ascribed to the considerably higher resubstitution cost in the classification problem. A comparison of Figs 1 and 2 and the prediction results (Figs 4 and 5 and Table 3) shows that the model with the more complex architecture (Fig. 2) hardly yields better modelling results. Obviously, regression problems and classification problems belong to two separate classes, whereby the parameters of the two types of models are different. Hence the above comparison can be an oversimplification. Nevertheless, similarly as in other regression methods, also in the case of the tree method, model simplicity should be a consideration in selecting a particular model.

4. Conclusions

There are many modelling methods, but recently the so-called machine learning methods, including the regression and classification tree method, are increasingly often used. At the present stage of the research on the operational reliability of municipal systems, modelling and predicting the failure rate and type of damage for water conduits seems to be critically important considering that in the case of serious failures, decisions must be promptly taken. As regards the use of the CT method for predicting the type of failure of water conduits, the results are not fully satisfactory because of the quite high percentage of incorrectly classified cases. This is rather disappointing since many of the actions taken by the network operator are connected with this type of failure. The approach to repairing a given section of a pipeline when the latter is fractured for a considerable distance along its length should be different than the one adopted to deal with a local corrosion pit or a perforation in the form of a hole. Therefore as part of further research some changes need to be made to the CT model in order to avoid so many incorrectly classified failure types. Perhaps when such variables as the pressure in the conduit, the temperature of the ground surrounding the pipeline and the pipeline laying depth are included in the vector of predictors, the quality of fit will improve.

In this paper RT models were applied to predict the failure rate of distribution pipes and house connections. For both DP and HC the RT models had one divided node and two end nodes. The resubstitution cost amounted to 0.0056 and 0.00073 for the model describing respectively the house connections and the distribution pipes. Even though only basic independent variables and a very simple tree architecture were used, the results are satisfactory, indicating that RT models can be an alternative to other modelling methods. Still, further research on RT models with the independent variables vector comprising variables previously not used in model building is needed. Such research, covering other water distribution systems, will soon be conducted to acquire data on the basis of which it will be possible to make some generalizations and advance further theses concerning the prediction of the failure rate and operational reliability of underground facilities.

As regards the CT model, the use of a more complex tree architecture than for the regression model did not translate into a higher quality of the predicted dependent variables (about 39% of incorrectly classified cases) or a lower cost (the resubstitution cost remained at about 0.39). Perhaps if two separate CT models are created for the distribution pipes and the house connections, as it was done in the case of the regression trees, the quality of the output (predicted) data will improve. This is another task for future research. In this paper for the first time an attempt has been made to apply the classification tree method to the modelling of types of water conduit failures in Poland. Further research in this area is both advisable and necessary.

Acknowledgement

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] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression tress. Boca Raton, USA: Chapman Hall/CRC; 1984. ISBN 978-0-412-04841-8.

[2] Statistica 13.1. Electronic Manual, https://www.statsoft.pl/textbook/stathome_stat.html? https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstsvm.html.

[3] Bevilacqua M, Braglia M, Montanari M. Reliab Eng Syst Safety. 2003;79(1):59-67. DOI: 10.1016/S0951-8320(02)00180-1.

[4] Rodriguez-Galiano V, Mendes M.P, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L. Sci Total Environ. 2014;476-477:189-206. DOI: 10.1016/j.scitotenv.2014.01.001.

[5] Sun H, Gui D, Yan B, Liu Y, Liao W, Zhu Y, et al. Energ Convers Manage. 2016;119:121-129. DOI: 10.1016/j.enconman.2016.04.051.

[6] Wang Z, Lai C, Chen X, Yang B, Shao S, Bai X. J Hydrol. 2015;527:1130-1141. DOI: 10.1016/j.jhydrol.2015.06.008.

[7] Malinowska A. Nat Hazards. 2014;73(2):317-334. DOI: 10.1007/s11069-014-1070-2.

[8] Li H, Sun J, Wu J. Expert Syst Appl. 2010;37(8):5895-5904. DOI: 10.1016/j.eswa.2010.02.016.

[9] Weng J, Zheng Y, Qu X, Yan X. Transport Res. 2015;57:30-41. DOI: 10.1016/j.trc.2015.06.003.

[10] Musz A, Kowalska B. Ecol Chem Eng S. 2015;22(2):219-229. DOI: 10.1515/eces-2015-0012.

[11] Pietrucha-Urbanik K, Studziński A. Ecol Chem Eng A. 2016;23(3):299-311. DOI: 10.2428/ecea.2016.23(3)25.

[12] Orłowska-Szostak M. E3S Web of Conferences. 2017;17:00069. DOI: 10.1051/e3sconf/20171700069.

[13] Kamiński K, Kamiński W, Mizerski T. Proc ECOpole. 2016;10(2):661-666. DOI: 0.2429/proc.2016.10(1)072.

[14] Iwanek M, Kowalski D, Kwietniewski M. Badania modelowe wypływu wody z podziemnego rurociągu podczas awarii [Model studies of a water outflow from an underground pipeline upon its failure]. Ochr Środ. 2015;37(4):23-26. http://www.os.not.pl/docs/czasopismo/2015/4-2015/Iwanek_4-2015.pdf.

[15] Kwietniewski M, Rak J. Niezawodność infrastruktury wodociągowej i kanalizacyjnej w Polsce [Reliability of water supply and wastewater disposal infrastructure in Poland]. Warszawa: Monographs of the Civil Engineering Committee at the Polish Academy of Sciences, Studies in Engineering No. 67; 2010. ISBN 978-8-389-68751-7.

[16] Tchórzewska-Cieślak B. Global Nest J. 2014;16(4):667-675. DOI: https://journal.gnest.org/sites/default/files/Submissions/gnest_01344/gnest_01344_published.pdf.

[17] Musz-Pomorska A, Iwanek M, Parafian K, Wójcik K. E3S Web of Conferences. 2017;17:00062. DOI: 10.1051/e3sconf/20171700062.

[18] Iwanek M, Suchorab P, Karpińska-Kiełbasa M. Periodica Polytechnica Civ Eng. Online first. Volume (2017), paper 9728. DOI: 10.3311/PPci.9728.

[19] Francis RA, Guikema SD, Henneman L. Reliab Eng Syst Safety. 2014;130:1-11. DOI: 10.1016/j.ress.2014.04.024.

[20] Kutyłowska M. E3S Web of Conferences.2017;22;00097. DOI: 10.1051/e3sconf/20172200097.

METODY REGRESYJNE DO PRZEWIDYWANIA POZIOMU AWARYJNOŚCI I RODZAJU USZKODZEŃ PRZEWODÓW WODOCIĄGOWYCH

Wydział Inżynierii Środowiska, Politechnika Wrocławska, Wrocław

Abstrakt: W pracy przedstawiono możliwość zastosowania drzew regresyjnych i klasyfikacyjnych (RT i CT) do przewidywania wskaźnika intensywności uszkodzeń przewodów wodociągowych oraz rodzaju uszkodzenia. Analiza wykorzystująca algorytm budowy drzew polega na znalezieniu zbioru logicznych warunków podziału oraz znalezieniu relacji pomiędzy predyktorami (zmiennymi niezależnymi) a zmienną zależną, co w konsekwencji prowadzi do uzyskania wyników prognozowania. Przewidywanie wskaźnika awaryjności przewodów rozdzielczych i przyłączy wykonano na podstawie danych eksploatacyjnych z lat 2008–2014 dla jednej wybranej strefy zasilania w wodę średniej wielkości polskiego miasta. Zmiennymi niezależnymi były: długość danego typu przewodów oraz liczba uszkodzeń zaobserwowanych w danym roku na rurociągach rozdzielczych i przyłączach. Stworzono oddzielne modele drzew regresyjnych do modelowania awaryjności przewodów rozdzielczych i przyłączy. W przypadku zagadnienia klasyfikującego zbudowano jeden model opisujący łącznie uszkodzenia zaobserwowane na rurociągach rozdzielczych i przyłączach. W tym modelu jakościową zmienną zależną był rodzaj uszkodzenia, a predyktorami materiał, średnica i typ przewodu. Uzyskane wyniki wskazują, że metoda RT może być stosowana do oceny poziomu awaryjności przewodów wodociągowych. Natomiast klasyfikacja rodzaju uszkodzeń nie była całkowicie satysfakcjonująca, co świadczy o konieczności prowadzenia dalszych badań w tym zakresie. Obliczenia przeprowadzono w programie Statistica 13.1.

Słowa kluczowe: metody regresyjne, sieć wodociągowa, rodzaj uszkodzeń przewodów wodociągowych