MARSplines method as a tool for failure frequency modelling

Małgorzata Kutyłowska[1,*]

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

Abstract. The paper presents the results of failure rate prediction using adaptive algorithm MARSplines. This method could be defined as segmental and multiple linear regression. The range of segments defines the range of applicability of that methodology. On the basis of operational data received from Water Utility two separate models were created for distribution pipes and house connections. The calculations were carried out in the programme Statistica 13.1. Maximal number of basis function was equalled to 30; so-called pruning was used. Interaction level equalled to 1, the penalty for adding basis function amounted to 2, and the threshold – 0.0005. GCV error equalled to 0.0018 and 0.0253 as well as 0.0738 and 0.1058 for distribution pipes and house connections in learning and prognosis process, respectively. The prediction results in validation step were not satisfactory in relation to distribution pipes, because constant value of failure rate was observed. Concerning house connections, the forecasting was slightly better, but still the overestimation seems to be unacceptable from engineering point of view.

1 Introduction

Water-pipe networks and other elements of water supply systems are known thousounds of years. Even in bronze epoch such municipal utilities were built. The first pressure pipes were exploited at the beginning of the second millenium before Christ in Minoan and Greek civilization [1]. Besides, also in ancient times water was used by people for recreation purposes, e.g. fountains and water reservoirs [2]. The basic hydraulic rules and the proper maintenance of water quality (using filtration and sedimentation processes) were also known in ancient times [1, 3].

In modern era the water treatment technology and water distribution has been widely spread in the most parts of Europe. The first pipes were made from wood. After some centuries clay, ceramic and stones were used as material for water conduits. Nowadays, the water supply systems are very important and essential parts of the whole municipal infractructure. Currently, problems of designing or building new sections of water-pipe networks are not so important, because most of households are connected to the buried infrastructure – water and sewerage systems. At present, more significant studies should be carried out in relation to proper maintainance of existing utilities. Exploitation and

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

maintenance of conduits are closely connected with failure and reliability analysis as well as with water losses and developing new approaches of calculations. Such inveatigations are done in Poland and abroad concerning water systems [4–7] and also sewers, especially storm-water systems which are now more vulnerable on some unpredictable weather conditions [8, 9].

Operational data used in failure analysis could be also very important in modeling processes. Nowadays, in Poland some studies are carried out just concerning mathematical approaches in risk or reliability analysis [10, 11]. That is the reason why also in relation to failure rate prediction some mathematical algorithms are taken into account. The main aim of this work is to check the possibility of using MARSplines method for forecasting of failure frequency of water pipes. This algorithm was recently used in geotechnical and chemical engineering [12, 13]. One attempt of using MARSplines method in environmental engineering was done by the author [14] and in the following paper sligthly different approach is proposed to verify hypothesis stated previously [14].

2 Methodology and range of studies

2.1 MARSplines method

MARSplines (*Multivariate Adaptive Regression Splines*) is a method using spline functions for prediction purposes [15]. Classification and regression problems could be solved by means of this algorithm. The significant advantage of this nonparametric method is that it is not necessary to know the relationships between dependent and independent variables (predictors). The function should not be defined *a priori*, as in other regression methods, e.g. neural networks or support vectors. The relation between two kinds of variables is based only on the analysis of set of factors and basis functions, which are selected from the modelling data. MARSplines method could be defined as segmental and multiple linear regression. The range of segments defines the range of applicability of that algorithm. Input data collected in the decision space are divided in such way to obtain separate subsets with specific regression or classification functions. This is some kind of advantage in the case when the vector size of input data (with a lot of variables) is relatively huge. Other predicting algorithms could be limited by the multidimensional problems. Mentioned above factors, defining the influence of the predictor on the dependent variable, could be compared by each other (for different independent variables) only in the case when the variables are normalized to 0 and 1 – average and standard deviation, respectively. Basis functions are linear (t-x) and (x-t). Parameter *t* is called node of basis function. Its value depends on the problem which is currently solved. Specific basis functions and model parameters (obtained by least square method) as well as input data are necessary to predict dependent variable. The general equation in MARSplines method could be defined as [15]:

$$y = f\left(X\right) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k\left(X\right) \tag{1}$$

The summing is done for all *K* model elements. Dependent variable *y* is calculated as the function of independent variables *X* (and their interactions). β_0 and β_k are called as initial and weighted ordinate of one or more basis functions h_k(X), respectively.

The cut basis functions for modelling of dependent variable are defined as [15]:

$$(x-t)=\begin{cases} x-t, & x>t \\ 0, & x\le t \end{cases} \tag{2}$$

MARSplines algorithm is used for exploring the decision space of input data and dependent variable as well as also for analysis of interactions between data. The aim of such approach is to maximize the level of fitting and finding the most significant (important) predictors. The risk of overtraining is possible, because MARSplines is nonparametric model and fits the data very well. Reduction of basis function (so-called pruning) is the way to avert the overfitting the predicted value to real one. The selection of the most important predictors is connected with the reduction of basis functions. Such function are selected, which have the great influence on the prediction of dependent variable [15]. Also the lowest increase of squared error is required. The model quality is described by GCV indicator (*Generalized Cross Validation*), which is approximation of cross error validation [15, 16]:

$$GCV=\frac{\sum_{i=1}^{N}\left(y_i-f(x_i)\right)^2}{N\left(1-\frac{C}{N}\right)^2} \tag{3}$$

$$C=1+cd \tag{4}$$

N – number of cases,
C – penalty for adding the next basis function,
d – effective of degrees of freedom; equals to the number of independent functions,
c – controls the penalty value.

2.2 Water-pipe network

The operating data registered in years 2001–2012 and received from water utility in one selected Polish city were used to predict failure rate (λ, fail./(km·a)) of distribution pipes and house connections. Operational data from the time span 2007–2012 were used for creating and building the MARSplines model for distribution pipes. On the basis of data from the previous 6 years (2001–2006) the prognosis (validation) of failure rate was carried out. For house connections the procedure was opposite. The first six years were taken into account in learning process, the next 6 years in validation step. Such approach was carried out intentionally to check if smaller size (in years 2007–2012) of vector with independent variables has an influence on the prognosis results. As it can be seen in the section 3 the number of training cases plays significant role in the learning quality and during creating the model.

Analyzed city is medium town in Poland with about 40000 inhabitants. Most of households (almost 100%) are connected to the water supply system. Water pressure in the pipe network amounted to ca. 0.3–0.4 MPa. The network consists of: grey cast iron and PE mains with diameters of DN 150 to 800 mm; distribution pipes with diameters of DN 50 to 315 mm, made variously of ductile or grey cast iron, steel, PE, PVC; house connections with diameters of DN 20 to 150 mm, made of steel and galvanized steel, grey cast iron, PE. In the analysed city mainly house connections are upgraded or replaced when they are characterized by a high failure frequency. The material structure in relation to the total length (at the end of 2012) of the main and distribution conduits is as follows: 64.1% – grey cast iron, 24.0% – PE, 7.4% – PVC, 2.0% – ductile cast, 2.5% – steel. The structure of the water-pipe network is displayed in the figure 1. The system expansion is especially observed in the suburbs in new housing developments.

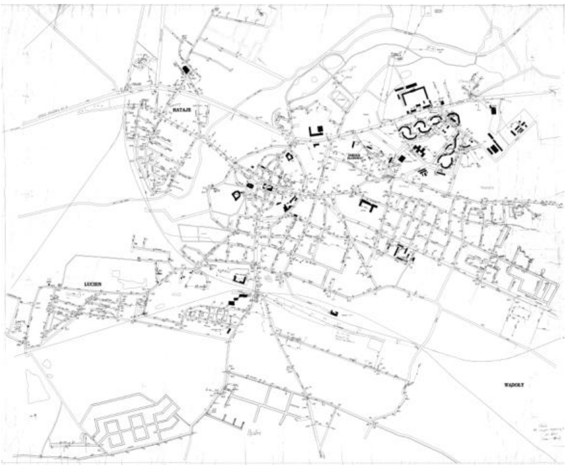

Fig. 1. Water-pipe network architecture [17].

The indicator λ was dependent variable which was forecasted using independent variables (predictors) like as: material, length, diameter and year of construction of each kind of conduits. For distribution pipes the ranges of each predictor in the time span 2007–2012 amounted to: length – 69.9–88.7 km, diameter – 90–200 mm, year of construction – 1961–2006, material – cast iron, steel, PE and PVC. Experimental dependent variable varied between 0.10 and 0.31 fail./(km·a). The ranges of each independent variable used in validation step were slightly different and equalled to: length – 57.3–59.7 km, diameter – 80–200 mm, year of construction – 1961–2006, material – cast iron, steel, PE and PVC. Real values of failure rate amounted to 0.34–0.57 fail./(km·a). For house connections the ranges of each predictor in the time span 2007–2012 amounted to: length – 31.4–50.2 km, diameter – 20–100 mm, year of construction – 1961–2009, material – cast iron, steel, galvanized steel and PE. Experimental dependent variable varied between 0.23 and 0.83 fail./(km·a). The ranges of each independent variable used in learning step were slightly different and equalled to: length – 23.4–29.9 km, diameter – 25–100 mm, year of construction – 1961–2006, material – cast iron, steel, galvanized steel and PE. Real values of failure rate amounted to 0.84–1.59 fail./(km·a).

The calculations were carried out in the programme Statistica 13.1. Maximal number of basis function was established at the level 30; unsufficient function was reduced (so-called pruning) during the process of builing the MARSplines model. Interaction level equalled to 1, the penalty for adding basis function was equal to 2, and the threshold – 0.0005. Interaction equals to 1 takes into consideration main influences between variables. The penalty is used when the next function is added to the algorithm. The threshold is used to minimize the possibility of overtraining and overfitting model results to real values.

3 Results and discussion

The main assumption made during the modelling process was that two separate models are built to predict failure rate of distribution pipes and house connections. The approach proposed in this paper is different than that in previous work [14], where one model was created to forecast indicators λ for water mains, distribution pipes and house connections. The results presented in the paper [14] were unsatisfactory from engineering point of view maybe just because one model with three dependent variables was responsible for predicting failure frequency. It is the reason why in the following work two models are proposed to check if that changes in the approach may have an influence on the prediction results or maybe MARSplines method is just not suitable for forecasting of failure rate of water pipes.

Analysis of the results obtained for distribution pipes shows us that the number of basis function was reduced from 30 to 1 during the modelling process. The higher number of basis function (in the case when the reduction was not applied) did not improve the correlation between experimental and forecasted values of failure rate. GCV error equalled to 0.0018 and 0.0253 in learning and prognosis process, respectively. The number of factors (ordinates) were equal to 2 defined as the initial ordinate (β_0 = 0.3091) and the length of conduits (β_1 = -0.0049). It means that only one predictor (length) was dominated and was the most responsible for failure rate prediction of distribution pipes. MARSplines method has one feature different that other regression algorithms. The equation describing the modelled problem is generated. The failure frequency of distribution pipes (λr) is described by the relation:

$$\lambda r = 0.3091 - 0.0049 \cdot max\left(0; length - 69.9\right) \tag{5}$$

The results of prediction of failure rate of distribution pipes are shown in the figure 2.

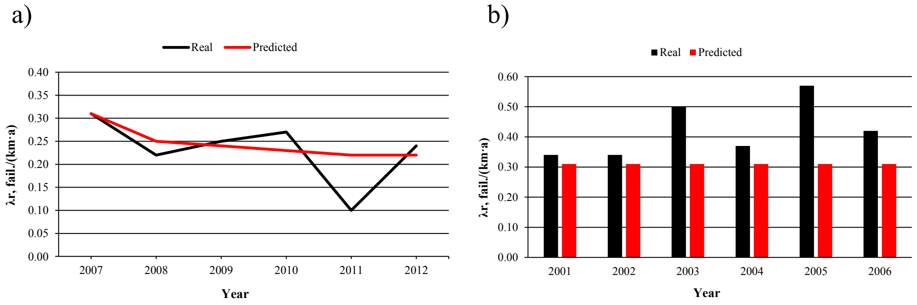

Fig. 2. Experimental and predicted failure rate of distribution pipes, a) learning step, b) validation step.

The correlation between real and forecasted values of failure rate (fig. 2) is not satisfactory. In the learning process (building the model – fig. 2a) the determination coefficient R^2 was equal to 0.38. It means that the convergence is rather poor. The decreasing tendency of predicted failure rate is not true because the value of indicator λr was fluctuating year by year. The results from validation process are even worse because the constant value (0.31 fail./(km·a) of predicted λr was observed. It means that MARSplines method could not be recommended as regression algorithm for forecasting of failure frequency of water pipes. Even building two separate models for each kind of conduit (different approach in comparison to previous studies [14]) did not change the lack of convergence between experimental and forecasted values of indicator λr. From all values of the length of this kind of conduit only one (69.9 km, in 2007) was selected by the model as the dominant one. It means that independent variables just from year 2007 played the significant role in the whole modelling process. This fact maybe had an influence on the prognosis results in the validation step (fig. 2b). The model has lost the generalization features and the proper prediction was not possible due to shortages in learning process.

For house connections the failure frequency prediction was described by another model than for distribution pipes. The number of basis function was equalled to 3 (the reduction from 30 using so-called pruning). The value of GCV error amounted to 0.0738 and 0.1058 in learning step and validation step, respectively. Although the value of GCV error was higher in relation to the model predicting *λp* in comparison to the model forecasting *λr*, the results of modelling and correlation between real and predicted failure rates are better concerning house connections (fig. 3). It means that the model quality should be assessed using not only the values of errors estimated during the model building, but also using: number of significant kinds of independent variables, the size of the learning vector and even experience of researcher. The most appropriate number of factors (in relation to failure rate of house connections) was equal to 4 and described by: β_0 = 1.1641, β_1 = -0.0162 (length), β_2 = 0.0067 (year of construction) and β_3 = -0.2569 (material-PE). It means that three independent variables should be treated as the most responsible for failure rate prediction. The most significant was especially year 1974 and pipe's length registered in 2005. What is more surprising is the fact that conduits made from plastics plays important role in the modeling of failure frequency of house connections. Only 7 cases from the whole independent variable vector (173 cases) in learning step were described as damage of polyethylene pipe. It means that for not known reasons model treated just this variable as really essential. MARSplines method could also belong to so-called „black box" approaches (as other regression algorithms). It means that sometimes it is difficult or even impossible to understand the choice done by the algorithms (in this case splines basis functions responsible for prediction) during the model creation. The failure rate *λp* is described by the equation:

$$\begin{aligned}\lambda p &= 1.31641 - 0.0162\,max\left(0;\,length - 23.4\right) + 0.0067\,max\left(0;\,year - 1974\right)\\ &- 0.2569\,max\left(0;\,material\left\{PE\right\} - 0\right)\end{aligned} \tag{6}$$

The effects of forecasting of indicator *λp* are displayed in the figure 3.

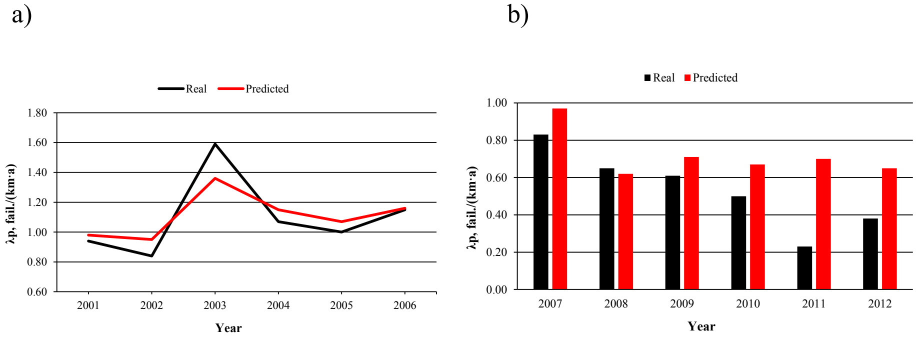

Fig. 3. Experimental and predicted failure rate of house connections, a) learning step, b) validation step.

The number of training cases was higher than in relation to distribution pipes which effects a little bit better prediction results in validation step. The values of indicator *λp* are in most cases slightly overestimated (Fig. 3) in both processes (learning and prognosis). In some years (e.g. 2010, 2011 and 2012) this overestimation is very high – even 3 and 2 times in comparison to experimental values of failure rate. For other years (beside these 3 indicated above) of investigations the correlation between real and forecasted values of indicator *λp* could be accepted from engineering point of view. The relative error varied between 0.87% and 14.47% (creating the model, R^2 = 0.94) as well as 4.62% and 16.87%

(validation step, R^2 = 0.36). In proposed in this paper approach two separated models for each kind of water conduit were created. It means that models are independent from each other and could be used for prediction purposes independently. Even if the results for distribution pipes are senseless, the model for house connections could be used for forecasting of failure rate, but some restrictions of this kind of modelling should be taken into account, e.g. the proper size of independent variable vector, the same (as in proposed model) kinds of predictors or even proper assessment of GCV error value.

4 Summary

MARSplines method was proposed for prediction of failure rate of water pipes. Two separated models for distribution pipes and house connections were created. The results of modelling for both kinds of conduit type show that this regression algorithm could not be rather recommended for forecasting of failure frequency. Although models are independent and the results of investigations in relation to house connections are slightly better than for distribution pipes still the determination coefficient in validation step is low (0.36). According to the short literature review, MARSplines method was not used earlier in environmental engineering in relation to failure analysis. The attempt described in this paper shows us that this kind of modelling is rather useless for failure rate prediction. The constant value of failure indicator in validation step concerning distribution pipes means that the methodology is senseless and could lead us to incorrect conclusions. Quite high overestimation of failure indicator of house connections also gives us improper response concerning the level of failure frequency. On the basis of modelling results e.g. renovation or replacing of some pipe sections could be performed. If the modelling solutions have relatively high errors, also the propositions of modernization way will be weighted with some inaccuracy. Probably MARSplines algorithm is too easy (using only splines basis functions) to forecast properly so complicated problem as level of failure of water-pipe network. It could be worth checking if in classification problems (e. g. classification of kinds of damages) MARSplines method gives better and more reasonable solutions.

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. H. Mala-Jetmarova, A. Barton, A. Bagirov, Wat. Sci. Technol. **15**, 2 (2015)
2. L.W. Mays, D. Koutsoyiannis, A.N. Abgelakis, Wat. Sci. Technol. **7**, 1 (2007)
3. L.W. Mays, Wat. Sci. Technol. **13**, 3 (2013)
4. C. Gong, W. Zhou, Struct. Infrastruct. E. **13**, 11 (2017)
5. B. Ward, A. Selby, S. Gee, D. Savic, Urban Water J. **14**, 7 (2017)
6. K. Pietrucha-Urbanik, A. Żelazko, Period. Polytech-Civ. **61**, 3 (2017)
7. A. Musz-Pomorska, M. Iwanek, K. Parafian, K. Wójcik, E3S Web of Conferences **17**, 00062, DOI: 10.1051/e3sconf/20171700062 (2017)
8. E. Kuliczkowska, Environ. Prot. Eng. **44**, 2 (2017)
9. M. Wdowikowski, A. Kotowski, P.B. Dąbek, B. Kaźmierczak, E3S Web of Conferences **17**, 00096, DOI: 10.1051/e3sconf/20171700096 (2017)
10. I. Zimoch, Ochr. Sr. **38**, 4 (2016)

11. B. Tchórzewska-Cieślak, Global Nest J. **16**, 4 (2014)
12. L.J. Wang, M. Guo, K. Sawada, J. Lin, J. Zhang, Catena, **135** (2015)
13. J. Antanasijević, V. Pocajt, D. Antanasijević, N. Trišović, K. Fodor-Csorba, Liq. Cryst. **43**, 8 (2016)
14. M. Kutyłowska, Application of MARSplines method for failure rate prediction, Conference Proceedings New Technologies in Water and Sewer Systems (to be published) (2018)
15. Statistica 13.1, Electronic Manual
16. Statsoft "Data mining – prediction methods", workshop materials, Kraków (2017)
17. Operational data received from Water Utility