Application of K-nearest neighbours method for water pipes failure frequency assessment

Małgorzata Kutyłowska[1,*]

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

Abstract. The paper describes the results of failure rate modeling using K-nearest neighbours method (KNN). This algorithm is one among other regression methods, called machine learning methods. The aim of the presented paper was to check the possibilities of application of such kind of modelling and the comparison between current results and investigations of failure rate prediction in another Polish city. Operational data from 12 years of exploitation, received from water utility, were used to predict dependent variable (failure rate). Data (249 and 294 for distribution pipes and house connections, respectively) from the time span 2001–2012 were used for creating the KNN models. On the basis of other data (one case for each year) the validation of optimal model, based on Euclidean distance metric with the number of nearest neighbours K = 2, was carried out. The realization of the modelling was performed in the software program Statistica 12.0.

1 Introduction

The assessment and analysis of failure frequency and reliability of water distribution systems and their elements have been very precisely described in the literature [1–3]. Nowadays, typical approach concerning the analysis of the number and kinds of occurring damages should be and is extended of the effects of water losses [4] which should be minimalized due to limited water resources in Poland. Moreover, more often the effects of failures are evaluated by consumers [5] and according to their opinion it is required to improve the quality of water delivering.

1.1 Modeling

The investigations related to the technical condition of buried infrastructure are extended and many mathematical modeling methods and special software are used to increase the efficiency of the management of water-pipe networks [6–8]. The prediction of failure rate could be carried out not only using typical statistical methods [9], but also using algorithms based on artificial intelligence [10–12]. Artificial intelligence known as neural networks based on unsupervised learning (Kohonen network) could be successfully used for risk analysis of water distribution system [13]. Till now K-nearest neighbours (KNN) method was not widely used for reliability assessment of municipal systems and prediction of selected indicators. The main aim of this work was to indicate the possibilities of using KNN algorithm for failure analysis of water conduits. This method was successfully applied in many scientific fields, e.g. for failure analysis of mechanical facilities [14] and in broadly understood medicine [15]. Hence it seems to be reasonable to check the usability of KNN method in failure frequency analysis of water supply systems.

1.2 K-nearest neighbours method

K-nearest neighbours algorithm is relatively easy in implementation and in analysis in comparison to other regression methods. KNN could be used in classification [16] or regression [17] problems. Modeling of failure rate of water pipes is based on regression not classification algorithm. The main assumption of this algorithm is to classify similar data to the same classes. The prediction of dependent variable is based on the comparison if this variable belongs to the exemplary set or not [18]. The choice of the number of K-nearest neighbours has the great influence on the model quality. The lower number of K, the bigger prediction errors. On the other hand, too many K-nearest neighbours can lead to modeling results with so-called error of overfitting. Similarly, as in other regression methodologies, it is required to find the optimal solution. Optimal number of K is not known *a priori*. The application of so-called V-fold cross validation is recommended for finding the most beneficial result in terms of the model quality and the agreement between dependent variable and experimental values. In this type of cross validation, data are divided into V randomly selected disjoint parts. Using the V-1 parts of data as training examples the dependent variable is predicted and the prediction error is calculated on the basis the residual sum of

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

squares. The procedure is executed for all the V data segments. Then a model quality measure is determined on the basis of the averaged errors of the particular cycles. The optimal model parameters are selected during a quality analysis [18]. In regression problems, the average for K nearest neighbours is calculated according to the equation (1) [18]:

$$y = \frac{1}{K}\sum_{i=1}^{N} y_i \quad (1)$$

where y_i is the output value for i learning example and y is the value of output variable for new example. The result is obtained on the base of the K nearest neighbours of new point. Following this assumption, it is needed to have some kind of measurement of the distance between examples. There are four types of distance metric: Euclidean (D_E) – equation (2), quadratic Euclidean (D_{E2}) – equation (3), Manhattan (D_M) – equation (4) and Czebyszew (D_C) – Equation (5) [18]:

$$D_E(x_n, p) = \sum_{i=1}^{N}\sqrt{(x_{ni} - p_i)^2} \quad (2)$$

$$D_{E2}(x_n, p) = \sum_{i=1}^{N}(x_{ni} - p_i)^2 \quad (3)$$

$$D_M(x_n, p) = |x_{ni} - p_i| \quad (4)$$

$$D_C(x_n, p) = \max_{1\le i\le N}(|x_{ni} - p_i|) \quad (5)$$

where $D(x_n, p)$ is the distance metric, x_n is the new point and p is the learning example. The regression or classification precision depends mainly on the metric used to calculate distances [19].

2 Methodology of studies

The calculations were carried out in the programme Statistica 12.0. Operating data, received from water utility, from the time span 2001–2012 were used for modeling separately (with different models) failure rates (λ, fail./(km·a)) of distribution pipes (DP) and house connections (HC). The whole data set (249 data for distribution pipes and 294 data for house connections) was randomly divided into two samples (learning – 75% and testing – 25%). Moreover, the model validation (verification) was carried out using data which did not belong to mentioned above two sets (one case for each year). KNN models were created on the basis of all four kinds of distance metrics (D_E, D_{E2}, D_M and D_C). Indicators (dependent variables) λ_r for distribution pipes and λ_p for house connections were predicted using independent variables. Independent variable (so-called predictors) vector contained (separately for distribution pipes and house connections): length-L, diameter-D and material-M (cast iron-Z, steel-S, galvanized steel-SO, PE, PVC) as well as the year of construction-YC of water pipes. The short description of investigated water-pipe network was presented in the work [20]. In the current analysis 10-fold cross validation method was used. Such approach was also applied in other regression algorithms, e.g. in regression trees. The experience of the precursor of regression tree methodology, Breiman et al. [21] indicated that just V = 10 is the optimal value. The range of experimental dependent and independent variables in years 2001–2012 is displayed in Table 1 (learning and testing sample). The number of failures of house connections as well as of distribution pipes in the learning sample varied respectively between 10 and 27 as well as 8 and 32 in the time span of 2001-2012.

Table 1. Dependent and independent variables – learning and testing.

	L, km	D, mm	YC	λ, fail./(km·a)
DP	57.3–88.7	80–200	1961–2006	0.10–0.57
HC	23.4–50.2	20–100	1961–2012	0.23–1.59

3 Results and discussion

In Fig. 1 and Fig. 2 the real (experimental) and predicted (by KNN model) failure rates of house connections and distribution pipes are displayed. The results of modeling concern the model validation step. Such approach seems to be reasonable, because the model quality assessment should be carried out using the data which were not included to the learning and testing step of modeling. For learning and testing step the prediction results were very good. Pearson correlation coefficient was established at the level of ca. 0.99.

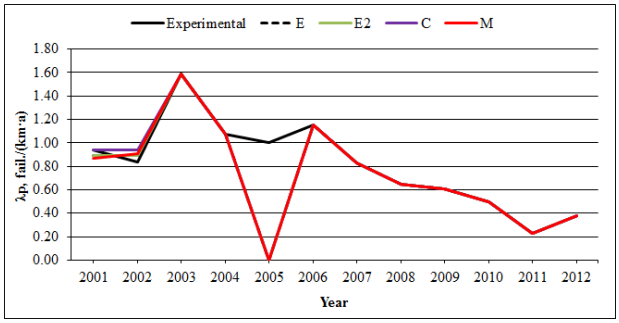


Fig. 1. Experimental and predicted failure rate of house connections.

The results of prediction of indicator λ for house connections (Fig. 1) are almost ideal convergent with real values for all kinds of distance metric. Some small differences between experimental and predicted failure rates (for all distance metrics) are visible for years 2001 and 2002. On the other hand, completely not understood is the result for 2005 when the model generated value of indicator λ equalled to zero for all distance metrics (D_E, D_{E2}, D_M and D_C). Operational data for year 2005 could

not be treated as outliers in comparison to other analysed years. It could be only assumption that modeling using K-nearest neighbours method contains the elements of so-called „black box" (similarly as other machine learning algorithms [22], which are also classified to regression methods). It means that deep inquire in the way of creating the optimal model and optimal results is rather impossible.

More surprising results of validation of the KNN model are observed for prediction of failure rate of distribution pipes (Fig. 2). In this case the ideal convergence between predicted and experimental values for each analysed year and all kinds of distance metric was achieved (in the Fig. 2 series KNN means that all distance metrics gave the same results).

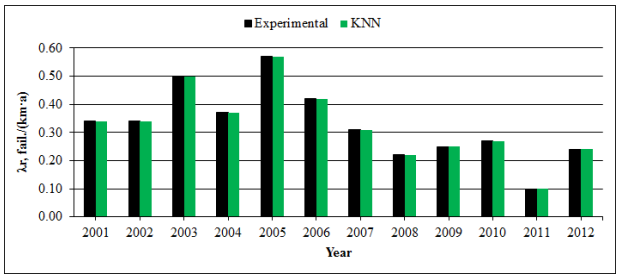
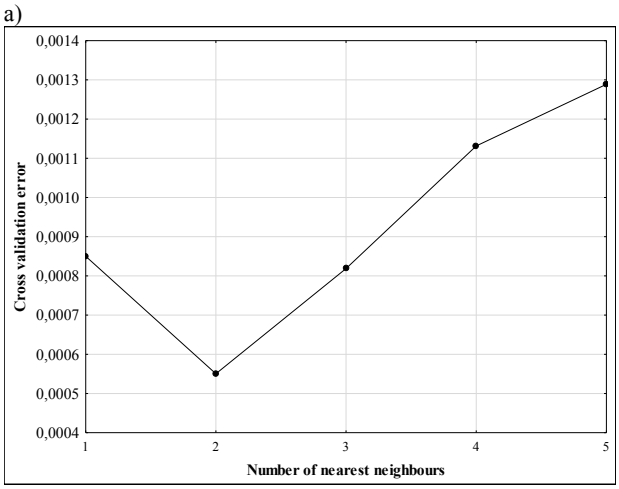


Fig. 2. Experimental and predicted failure rate of distribution pipes.

Completely different results were observed during the failure analysis also using KNN model on the basis of operational data from another Polish city [23]. It is obvious that each kind of modelling and forecasting is and even should be burden by the prognosis error which results from not only the data inaccuracy, but also from the difficulties in describing the relations between dependent (λ) and predictors. Hence it is required to be careful when such ideal results as above are observed. Ideal agreement between modeling results and real values should raise doubts due to many limitations of mathematical modeling connected with the character of changing dynamically the natural phenomena describing the technical condition and failure frequency of water distribution system. Maybe the differences in modeling results displayed in this work in comparison to solutions obtained earlier [23] are connected with the kinds of predictors as well as with the size of the independent variable vector. The analyses of the results in both investigations indicate that the kind and number of independent variables have the great influence on the modeling quality. This aspect is maybe even more important than the number of neighbours K or the kind of distance metric. Moreover, there is one more difference in the approach presented in this work in comparison to previous modeling attempts by means of KNN [23]. The kind of the created model is significant issue. In this work two separate models for failure rate prediction of distribution pipes and house connections were built. In contrast, in another author investigations [23] one model was analysed. This one model had three dependent variables: indicators not only λ_p, λ_r (for distribution pipes and house connections) but also indicator λ_m (water mains). Despite the fact that in currently analysed case different independent variables, which in better way describe the water conduits, were used, created models seem to be simpler than model with three output variables. This fact is obviously connected with the modeling quality. More complicated models (with more complicated structure) could be burden by higher prediction errors than models with different predictors but with one dependent variable in the output.

The changes of cross validation errors values in the contrast with the number of nearest neighbours are displayed in the Fig. 3. The optimal number of K is also indicated.

a)



b)

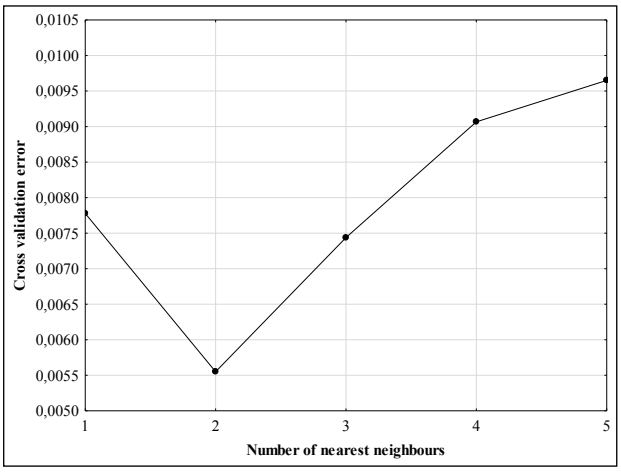


Fig. 3. Changes of cross validation error vs. number of K, a) distribution pipes, b) house connections.

Concerning distribution pipes and house connections, the optimal number of nearest neighbours is the same and equals K = 2. The maximum number of nearest neighbours (5), determined during the model creation, depends on the number of independent variables and on the number of learning data. The analysis of the figure 3 indicates that the lowest value of the cross validation terror (0.00555 – house connections and 0.00055 – distribution pipes) was observed in the models described by Euclidean (D_E) distance metric. Moreover, when the minimal values were achieved, the error value

was increasing with the increase of the number of K. Despite the fact that the prediction results (Fig. 1 and 2) do not differ in relation to distance metric, it seems that Euclidean distance metric is the optimal one for describing the failure rates of house connection as well as for distribution pipes. Application of other distance metric had the great influence on the higher number of K, what meant that the complexity of the model was bigger. The architecture (structure) of the model should be as simple as possible. In addition in many other modeling, by means of KNN method, problems just Euclidean metric was proposed as the optimal one [19]. This distance metric describes relatively good the relations between dependent variables and predictors, especially when this dependence is not known *a priori* [19].

4 Conclusions

The main aim of this work was to verify if application of K-nearest neighbours method is possible for prediction of failure rates of water pipes. The modeling, on the basis of operational data from the time span 2001–2012 received from one water utility, of the indicator λ (separately for distribution pipes and house connections) was performed using following independent variables: material, diameter, length and year of construction of water conduits. Models were created based on all kinds of distance metric (Euclidean, quadratic Euclidean Manhattan and Czebyszew). The maximum number of nearest neighbours was established at the level of $K = 5$. The increase of the number of K leads to increase of 10-fold cross validation error. The validation results of optimal models based on Euclidean distance metric with $K = 2$ are satisfactory. The ideal convergence between predicted and experimental values of failure rate was obtained. On the other hand, one should be very careful and sceptical when such great correlation is observed, because this fact could denote that model was overtraining. In such case the predicted values are ideally fit to learning sample what could mean that the generalization capabilities are lost. On the basis of the results shown in this paper one can conclude that KNN method seems to be less appropriable for failure rate prediction of water-pipe network than other regression algorithms (so-called learning machine methods), e.g. artificial neural networks, regression tress and support vector machine. The modeling does not mean the ideal agreement between the experimental and predicted values of variables, but rather means the recognition of the relationships between predictors and dependent variable. Many studies conducted until now in the whole world show that KNN method has not been applied for analysis and the assessment of failure frequency of water conduits. That is the reason to use this algorithm in the failure analysis of water-pipe network in Poland. It should be indicated that the choice of number and kinds of independent variables as well as the whole modeling process were carried out with the knowledge about KNN in relation to other engineering issues. In such case it is necessary to deepen the modeling methodology in relations exactly to investigations connected directly with reliability analysis of municipal systems. Maybe, this hypothesis should be checked using operational data from other water distribution systems, KNN method, in distinction to other regression algorithms, does not require too large (number of cases and variables) vector of predictors. Generalization ability maybe has been just lost due to using ca. 200 learning cases. If such hypothesis is confirmed during the modeling of failure rate of water pipes in another city, it will mean that KNN algorithm is privileged to other regression methods on account of difficulties of obtaining huge operating data.

The work was realized within the allocation No. 0401/0069/16 awarded for Faculty of Environmental Engineering Wroclaw University of Science and Technology by Ministry of Science and Higher Education in years 2016–2017.

References

1. H. Hotloś, *Quantitative assessment of the effect of some factors on the parameters and operating costs of water-pipe networks* (Wrocław University of Technology Publishing House, Wrocław, 2007)
2. M. Kwietniewski, J. Rak, *Reliability of water supply and wastewater infrastructure in Poland* (Polish Academy of Science, Warszawa, 2010)
3. I. Zimoch, Ochr. Sr. **34**, 4 (2012)
4. M. Iwanek, A. Musz, B. Kowalska, D. Kowalski, M. Chołody, Instal **1** (2016)
5. J. Rak, K. Pietrucha-Urbanik, Instal **2** (2016)
6. B. Tchórzewska-Cieślak, D. Szpak, Ochr. Sr. **37**, 3 (2015)
7. A. Scheidegger, J.P. Leitao, L. Scholten, Water Res. **83** (2015)
8. I. Zimoch, Ochr. Sr. **30**, 3 (2008)
9. Y. Kleiner, B. Rajani, Urban Water **3**, 3 (2001)
10. M. Kutyłowska, Instal **1** (2016)
11. M. Kutyłowska, Period. Polytech-Civ. **61**, 1 (2017)
12. M. Nishiyama, Y. Filion, Can. J. Civil Eng. **41**, 10 (2014)
13. K. Pietrucha-Urbanik, B. Tchórzewska-Cieślak, *Water Supply System operation regarding consumer safety using Kohonen neural network* (in Safety, Reliability And Risk Analysis: Beyond The Horizon, 2014)
14. A.B. Andre, E. Beltrame, J. Wainer, Appl. Artif. Intell. **27**, 1 (2013)
15. H. He, W. Graco, X. Yao, *Application of genetic algorithm and K-nearest neighbour method in medical fraud detection* (in SEAL'98, Springer-Verlag, 1999)
16. C.H. Wana, L.H. Lee, R. Rajkumar, D. Isa, Expert Syst. Appl. **39**, 15 (2012)
17. A. Mangaro, F. Pizzo, A. Lombardo, A. Pogliaghi, E. Benfenati, Chemosphere, **144** (2016)

18. Statistica 12.0, Electronic Manual
19. K.Q. Weinberger, L.K. Saul, J. Mach. Learn Res. **10** (2009)
20. M. Kutyłowska, M. Orłowska-Szostak, Water Practice and Tech. **11**, 1 (2016)
21. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and regression tress* (Chapman&Hall/CRC, Boca Raton, 1984)
22. S.H. Hamraz, S.S. Feyzabadi, *General-purpose learning machine using K-nearest neighbors algorithm* (in RoboCup 2005, Springer-Verlag, 2006)
23. M. Kutyłowska, *Prediction of failure rate of water pipes using K-nearest neighbours method* (in IWA 8th Eastern European Young Water Professionals Conference, 2016)