K-Nearest Neighbours Method as a Tool for Failure Rate Prediction

Małgorzata Kutyłowska

Abstract
The paper shows the results of failure rate prediction using non-parametric regression algorithm K-nearest neighbours. The whole data set for years 1999-2013 was divided randomly into two groups (learning – 75% and testing – 25%). Besides, data from year 2014 were used for verifying the model. The dependent variable (failure rate) was forecasted on the basis of independent variables (number of installed house connections, total length and number of damages of water mains, distribution pipes and house connections). Four types of distance metric: Euclidean, quadratic Euclidean, Manhattan and Czebyszew were checked and four KNN models were created. Taking into consideration all constraints and assumptions, models using Euclidean and quadratic Euclidean distance metrics gave the most optimal prediction results. The optimal number of K nearest neighbours equalled to 2 and 3 concerning models KNN-E, KNN-E2, KNN-C and KNN-M, respectively. Validation error was the smallest for models KNN-E and KNN-E2 and amounted to 0.0130, for model KNN-M was equal to 0.0152 and for KNN-C to 0.0150.

Keywords
collapse analysis, K-nearest neighbours, prediction, water-pipe network

1 Introduction
Water-pipe networks are one of the most important parts of the whole water supply system and belong to the critical infrastructure. The technical condition of the water conduits and amount of water provided to the consumers [1] should be maintained at the proper level concerning forecasting, suitable management and failure analysis. Research devoted to the determination of the water conduit failure intensity indicator and other factors having a bearing on the proper functioning of municipal water networks (e.g. impact of water losses on the soil surrounding the pipe [2]) has been conducted in Poland and in the world for many years [3], [4], [5], [6]. Failure rate of water pipes (λ) should be estimated not only on the basis of operational data, but also using the best available mathematical techniques and models, e.g. [7], [8]. On the other hand, failure analysis could be established using models based on the artificial intelligence [9], [10] or other statistical and probabilistic methods used in water and sewerage systems [11], [12], [13].

Recently, a lot of regression methods (e.g. support vector machine – SVM, regression trees – RT and K-nearest neighbours – KNN) were used to solve many engineering problems. For instance, the localization of leakages from water-pipe network was estimated using SVM algorithm [14], the building damage was assessed by means of RT [15] and KNN algorithm was used relating to time series analysis in industrial processes [16]. The main aim of this paper is to check if non-parametric regression algorithm KNN could be also useful for prediction of indicator λ of water conduits (water mains, distribution pipes and house connections).

2 Material and methods
K-nearest neighbours algorithm is the relatively simple one among other learning methodologies. It is assumed that similar data are grouped to the same class. The prediction is based on the comparison whether forecasted values belong to the exemplary set or not [17]. In the regression problems continuous dependent variable is predicted on the base of independent variables. The choice of the number of K nearest neighbours has significant meaning. This parameter is the most important concerning the
prediction quality. The smallest $K$, the bigger prediction variance. The optimal number of $K$ is not known a priori and usage of V-fold-cross-validation algorithm is recommended to find the best $K$. The main idea of cross-validation method is based on such approach: the data are divided into $V$ (chosen randomly) separate parts. The analysis is carried out for certain values of parameters using $V$-1 data sets as learning examples. In regression problems the prediction error is calculated as sum of squares of residuals. The procedure is repeated for all $V$ data segments and at the end the errors are averaged. Cross-validation algorithm is related to the estimation of prognostic quality of the model using testing sample which was not known for the model during its creation. In other words, the model is created using learning sample and the real model accuracy is checked using testing sample [17]. After selecting the proper number of $K$, the prediction could be carried out. In regression problems, the average for $K$ nearest neighbors is calculated according to the equation (1) [17]:

$$y = \frac{1}{K} \sum_{i=1}^{K} y_i$$

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 (E) – equation (2), quadratic Euclidean (E2) – equation (3), Manhattan (M) – equation (4) and Czebyszew (C) – Equation (5) [17]:

$$D(x, p) = \sqrt{(x - p)^2}$$

$$D(x, p) = (x - p)^2$$

$$D(x, p) = \text{Abs}(x - p)$$

$$D(x, p) = \text{Max}(|x - p|)$$

where $D(x, p)$ is the distance metric, $x$ is the new point and $p$ is the learning example. The regression or classification precision depends mainly on the metric used to calculate distances [18].

The calculations were performed in the programme Statistica 12.0. Operating data from the time span 1999–2014 (received from Water Utility) in one Polish water-pipe network were used for prediction purposes. The whole data set for years 1999-2013 was divided randomly into two groups (learning – 11 years and testing – 4 years) sample. The prediction results displayed in the tables 2–4 are related to testing sample. The learning sample was treated only as an example and the prediction was not carried out. The prediction results using verification sample (one year) are shown in the Figures 1–3.
The analysis of failure rate prediction of water mains (Table 2) indicates that models using Euclidean and quadratic Euclidean distance metrics are the most optimal. The convergence between experimental and predicted values of indicator $\lambda$ is relatively good.

### Table 2 Experimental and predicted failure rate of water mains (testing)

|          | $\lambda_m$, fail./(km·a) | KNN-E | KNN-M | KNN-C |
|----------|---------------------------|-------|-------|-------|
| Experimental | 0.32                       | 0.32  | 0.27  | 0.22  |
|           | 0.14                       | 0.18  | 0.14  | 0.13  |
|           | 0.14                       | 0.13  | 0.14  | 0.13  |
|           | 0.17                       | 0.13  | 0.13  | 0.13  |

On the other hand the best prediction of $\lambda_m$ in verification step (Fig. 1) was obtained using Czebyszew distance metric. KNN models (relating to different distance metrics) were created all together for water mains, distribution pipes and house connections. In other words one model consisting of all independent variables $L_P$, $L_m$, $L_r$, $L_p$, and $N_m$, $N_r$, $N_p$ was responsible for forecasting three dependent variables $\lambda_m$, $\lambda_r$, and $\lambda_p$. Taking into consideration this approach, it seems to be reasonable to choose such model which would generate the smallest discrepancies between experimental and predicted values of failure rate relating to three types of conduits.

The quality of the prediction is also measured by the error of the validation. This error was equal to 0.0130, 0.0152 and 0.0150 for models KNN-E and KNN-E2, KNN-M and KNN-C, respectively. It means that models using Euclidean and quadratic Euclidean distance metrics seem to be the most optimal.

### Table 3 Experimental and predicted failure rate of distribution pipes (testing)

|          | $\lambda_r$, fail./(km·a) | KNN-E | KNN-M | KNN-C |
|----------|---------------------------|-------|-------|-------|
| Experimental | 0.26                       | 0.38  | 0.38  | 0.36  |
|           | 0.35                       | 0.35  | 0.34  | 0.33  |
|           | 0.24                       | 0.33  | 0.34  | 0.33  |
|           | 0.42                       | 0.33  | 0.29  | 0.33  |

Concerning the failure rate prediction of distribution pipes it is obvious that not only in testing (Table 3), but also in verification (Fig. 2) KNN-E and KNN-E2 models are characterized by relatively good agreement between experimental and predicted values. For all types of conduits model KNN-C generated the constant value of indicator $\lambda$ for three, from among four, testing years. It means that the model using Czebyszew distance metric is rather not recommended for forecasting purposes. Moreover, the distance is measured as a maximum of absolute value of differences between new and example points (equation (4)). Probably such kind of measurement of distance between points is not suitable for prediction of failure rate of water pipes.

### Table 4 Experimental and predicted failure rate of house connections (testing)

|          | $\lambda_p$, fail./(km·a) | KNN-E | KNN-M | KNN-C |
|----------|---------------------------|-------|-------|-------|
| Experimental | 0.76                       | 0.75  | 0.70  | 0.65  |
|           | 0.53                       | 0.47  | 0.43  | 0.35  |
|           | 0.36                       | 0.35  | 0.43  | 0.35  |
|           | 0.64                       | 0.35  | 0.57  | 0.35  |

From engineering point of view the prediction results (Table 4), using models KNN-E and KNN-E2, relating to house connections are satisfactory and could be accepted.

The analysis of verification results (Fig. 3) indicates that model KNN-C forecasted indicator $\lambda_p$ the most properly in comparison to other models. But taking into consideration the constraints and assumption that the model should be suitable for all types of conduits, also in testing step, model KNN-C does not seem to be the optimal one.
The graph of the changes of cross-validation error depending on the number of nearest neighbours is displayed in the Figure 4. The maximum number of nearest neighbours was estimated at the level of 11. This number was established by the algorithm in the programme Statistica 12.0.

Maximal number of nearest neighbours depends on the number of independent variables and number of cases. The analysis of the Figure 4 shows that the lowest validation errors were obtained when the optimal (minimum) number of K-nearest neighbours was equal to 2 and 3 for models KNN-E, KNN-E2, KNN-C and KNN-M, respectively. After the errors reached the lowest value, then the errors were increasing. We can assume that even if the number of nearest neighbours is higher, the errors will be still increasing.

**4 Conclusions**

The K-nearest neighbours algorithm could be used for prediction of failure rate of water pipes. Analysis of the results obtained in the testing step shows that Euclidean and quadratic Euclidean distance metric seem to be the most appropriate for failure rate prediction for all types of pipelines. On the other hand, in the verification step Czebyszew and Euclidean distance metrics were the most suitable for failure rate forecasting of water mains, house connections and distribution pipes, respectively. Taking into account all assumptions and constraints, models KNN-E and KNN-E2 were supposed to be chosen as optimal. In this paper failure rate of water mains, distribution pipes and house connections were forecasted using independent variables related to three types of conduits. It means that each model consisted of three dependent variables \(\lambda_m, \lambda_r, \lambda_p\). Such assumption made at the very beginning forced to choose one model which was responsible for proper prediction of indicator \(\lambda\) of each type of water pipe. The results of forecasting and also the analysis of the lowest validation error pointed that the models characterized by Euclidean and quadratic Euclidean distance metrics were optimal. They also had the lowest number of \(K\) nearest neighbours which was equal to 2. It means that the models KNN-E and KNN-E2 were relatively simple that is quite important because among other aims the simplicity of the model is crucial issue in modelling.

The paper shows the very beginning step of failure rate modelling using K-nearest neighbours algorithm. The investigations presented in this paper will be expanded. The next step of researches would be checking whether other independent variables like e.g. material, diameter and year of installation of the water pipes are significant one for prediction purposes. The
models, for indicator $\lambda$ separately for each type of conduit, will be also created and checked whether the convergence between experimental and forecasted values is better than concerning one model as it was presented in this paper.

References

[1] Cieżak, W., Cieżak, J. "Routine forecasting of the daily profiles of hourly water distribution in cities. An effectiveness analysis". Environment Protection Engineering, 41(2), pp. 179–186. 2015. https://doi.org/10.5277/epci150215

[2] Iwanek, M., Suchorab, P., Karpinska-Kieblasa, M. "Suffosion holes as the result of a breakage of a buried water pipe". Periodica Polytechnica Civil Engineering, 61(4), pp. 700–705. 2017. https://doi.org/10.3311/PPci.9728

[3] Martinez-Codina, A., Castillo, M., Gonzales-Zeas, D., Garrote, L. "Pressure as a predictor of occurrence of pipe breaks in water distribution networks". Urban Water Journal, 13(7), pp. 676–686. 2016. https://doi.org/10.1080/1573062X.2015.1024687

[4] Kutyłowska, M. "Neural network approach for failure rate prediction". Engineering Failure Analysis, 41, pp. 23–29. 2014. https://doi.org/10.1016/j.engfailanal.2013.07.019