DEVELOPING GENETIC PROGAMMING AND NONLINEARREGRESSION ANALISYS MODELS FOR TURBIDITY OF DRINKING WATER

Bahrudin Hrnjica, Ali Danandeh Mer, Samira Dedić
University of Bihać,Technical faculty, Department of Mechanical Engineering,77000 Bihać,
bahrudin.hrnjica@unbi.ba,
Near East University, Department of Civil Engineering,Nicosia, North Cyprus,
ali.danandeh@neu.edu.tr
University of Bihać, Biotechnical faculty, 77000 Bihać, samira.dedic@yahoo.com.

Keywords: Genetic Programming, Regression Analysis, GPdotNET, Turbidity, Drinking Water.

SUMMARY:

Genetic programming (GP) and Regression analysis were applied to develop predictive modelsfor turbidity of drinking water in the main water source of Bihać town,Bosnia and Herzegovina. Predictive modelswere built based on monthly measurements data at the period 2006-2016. Turbidity measurement was collected based on the following parameters:KMnO4 consumption, daily rainfall, NH4+and period of passed time. All collected data was split in to training and testing data sets. Training data set was used to build models, while testing data set was used to control reliability and overfitting properties of the models.GP model was evolved using GPdotNET computer program. The result of modelling showed very high performance (R^2>0.85) of all models. Comparison analysis has shown that regression models are not reliable for predictionfor such nonlinear process, since performance parameters have lower values than GP model. Prediction for the next 12 months showed better performance and accuracy for the model build using genetic programming in comparison with the model given by regression analysis.

1. INTRODUCTION

One of the main quality parameter for drinking water is turbidity. It can be defined as cloudiness of a fluid which is caused by large number of invisible to the naked eye particles. Turbidity can be seen similar as the smoke in the air. We may say the turbidity is the key of water quality. Many studied prove the higher level of turbidity in drinking water would cause risk that people can develop gastrointestinal diseases [1]. High level of turbidity in drinking water can cause problems for immunocompromised people due to suspended solids. Suspend solids can contain various types of viruses or bacteria and develop disease while consuming this kind of water. To protect people, governments have set standards for allowable turbidity in drinking water. In Europe, each country developed its own standard which is based on the European standards for turbidity ISO 7027-1:2016. The standard specifies two quantitative methods using optical turbidimeters or nephelometers for the determination of turbidity of drinking water [2]:

- nephelometry, procedure for measurement of diffuse radiation, applicable to water of low turbidity (for example drinking water);
- turbidimetry, procedure for measurement of the attenuation of a radiant flux, more applicable to highly turbid waters.

Bosnia and Herzegovina has specific structure of soil which affect the turbidity, and it was necessary to develop its own standard. The standard which provides all procedures for maintain turbidity in drinking water is BAS EN ISO 7027:2002 which is based on the mentioned ISO standard. Usually, the turbidity is measured in nephelometric turbidity units (NTU). Different standards define different allowable levels of NTU. For example,value of turbidity in the drinking water is different. InUSA, allowable value of turbidity must be less than 5 NTU [4], while in European standards, the maximum values must not be more than 4 NTU [3].

The main causes of sudden appearance of turbidity in water is connecting with short-term appearance of rainfall by causing erosion of soil [5]. However, researchers have shown that turbidity development is a bit more complex process[6], and mainly depends of the soil structure, level of underground water, and variety of natural and anthropogenic factors and processes. Due to the fact the turbidity is very important parameter and directly influences the quality of drinking water, engineers used different approach to predict it[7]. In this paper, the turbidity modelswere developed based on Genetic programming method which was approved to be effective method in modelling and predicting water parameters [7, 11].

2. GENETIC PROGRAMMING

The use of artificial intelligence methods to solve engineering problems was intensified by the rapid development of information technology at the beginning of the 1990s of the last century. As hardware components start to be more powerful every year, methods of artificial intelligence and machine learning gets more and more powerful. In early 90s the first paper about genetic programming methods published [8]. Since the beginning genetic programming was inspired many engineers and scientist to apply method in different fields.

Genetic programming (GP) evolved as a generalization of 40 years old genetic algorithm (GA). In GA, chromosomes are representedmostly as binary numbers. With similar analogy, it is possible to create chromosomes which represent computer programs, as potential solution to problem [11]. Chromosomes in GP are represented in the population shaped like a hierarchical structure. They are constructed of primitive functions and terminals. Set of primitive functions can be any arithmetic operations, mathematical functions, Boolean operators, and special functions.Terminal set is also part of chromosome structure, and it is usually formed from input parameters and numerical constants. GP population, which consists of chromosomes, breads in consistence with natural principle of survival of the fittest, through genetic crossover, mutation and reproduction operations adopted to mating of computer programs.

GP begins with initialization of initial population of randomly generated computer programs. Each computer program (chromosome) in the population is evaluated by its ability to solve the problem, so called fitness function. Population is evaluated by computing the fitness value for each chromosome, which give the ability to select the best possible chromosome in the population. Beside the fitness function, genetic programming algorithm can be controlled with 19 parameters, of which there are 2 basic parameters, 11 secondary parameters and 6 qualitative variables that are selected with different alternative ways of executing the algorithm[8].

Two main parameters are:

- $\boldsymbol{M}$ - population size, and
- $\boldsymbol{G}$ – maximum number of generations.

Secondary parameters, a genetic programming are about probability value that a certain event occurs, which can be a specific genetic operation, i.e. quantitative value indicating the depth of s-expressions are:

- p_c – the probability of crossover; the recommended value should be greater than 90%,
- p_r – the probability of reproduction; it is recommended that it should be about 20% of the total population,
- p_{ip} – the probability of selection of crossover interior point (the functional node). It is recommended that the value is 90% probability to choose interior point for the crossover, while the analogous 10% would be the selection of external point (terminal node) for the crossover in relation to the total points of one chromosome,
- d_{formed} - maximum depth of S-expressions, formed on the basis of genetic operations of chromosome crossover or any other secondary genetic operation (mutation, decimation, encapsulation, etc.),
- $d_{initial}$ – maximum depth of S-expressions at the formation of the initial population,
- p_m – probability of mutation in the population,
- p_p – probability of permutation in the population,
- f_{ed} – frequency of decimation applications on the chromosome in the population,
- p_{en} – the probability of encapsulation in the population,
- u_{dc} – a condition for the activation of decimation on chromosomes,
- p_d – the percentage of the decimation in the population.

Secondary parameters are included in the algorithm depending on the method of GP implementation and are considered optional.

3. EXPERIMENTAL RESEARCH

In the studyGP was used to build mathematical model for turbidity of drinking water. Samples of water was taken at Klokot location – the main water supply source of Bihać town.

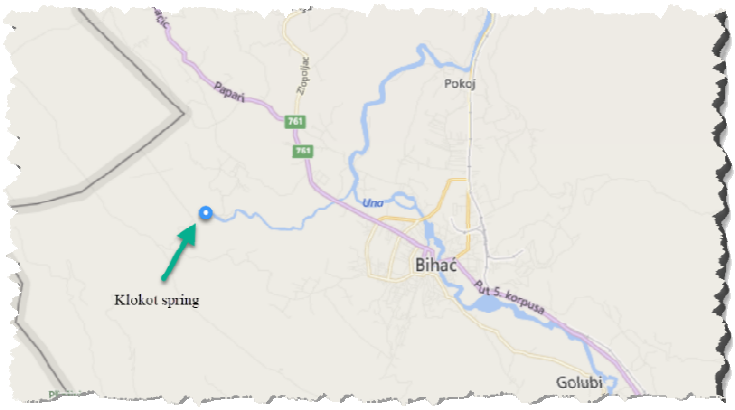

Figure 1. Klokot spring location

The turbidity of drinking water was measured with several input parameters: KMnO4 consumption, daily rainfall, NH4+ and period of passed time. Measures represent monthly data set which was collected for the period at 2006-2016.All collected data was split in to training and testing data sets. Training data set was used to build amodel, while testing data set was used to validate themodeloverfitting and prediction of turbidity for the next 6 months. In order toevaluate the model, nonlinear regression analysis was performed on the same data, and regression coefficients were calculated for several regression models. Furthermore, it was performed model evaluation in order to

get how the model determined with genetic programming is better than regression models of any level. GPdotNET [9] computer program was used for modelling and determination ofGP model.Wolfram Mathematica [10] was used to get regression models and comparison analysis. The following graph shows measured data:

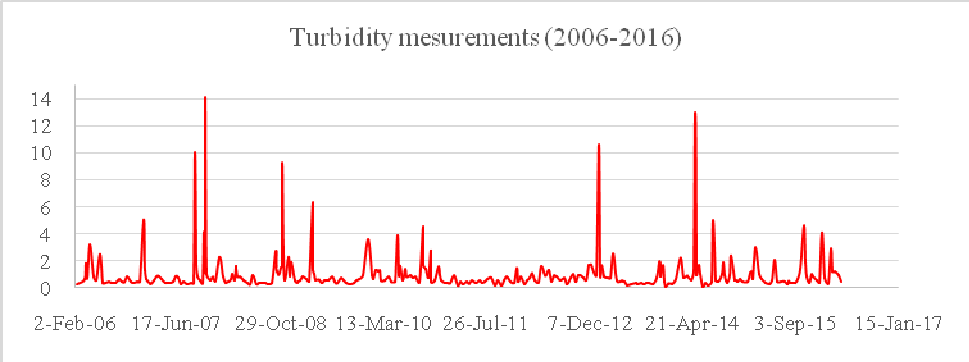

Figure 2: Turbidity of drinking water mesured at Klokot spring

Figure 2 shows turbidity values measured from 2006 to 2016 with respect to input parameters KMnO4 consumption, Daily Rainfall, NH4+ and period of passed time.

3.1 Building GP model for turbidity of drinking water

GPdotNETis used in order to get GP model of turbidity. During modelling the following GP parameters were used:

- Function set $F = \{+, -, *, /, 1/x, tanh\}$,
- Terminal set $T = \{X1, X2, X3, X4, \Re\}$,where$T_i$ – input parameters,$\Re$ - set of random constants.
- $M = 1000$,
- $G = 500$,
- $p_c = 95\%, p_m = 20\% p_r = 10\%, d_{initial} = 5, d_{formed} = 9$.

Before the searching process is started, training and testing data must be loaded, GP parameters are set and termination criteria was setup to 500 generations (iterations). At the beginning of the searching process, several different configurations were applied in order to get the best possible combination of all GP parameters. The time execution of the searching process in GP depends on the complexity of the process and the accuracy to be achieved. When a satisfactory mathematical model is acquired, post-processing of the results is enabled, where the mathematical model is exported to other programs for the evaluation and comparison with other results. GP model of turbidity is given in analytical form in the following expression.

$$y(x_1, x_2, x_3, x_4) = 1.29 + 15.39\left(0.107 + \left(0.156\,(-0.51 + x_1) + 0.007\,x_2 x_3(-0.015 + 0.029\,x_1 + x_4)\sqrt[3]{0.476 + 0.062\,x_2 - 4.073\,x_4^2}\right)\mathrm{Tan}(2.207\,x_4)\right). \quad (1)$$

where x_1 - KMnO4 consumption, x_2 - Daily Rainfall, x_3- NH4 and x_4- period of passed time.

Calculated turbidity values are shown on the Figure 3.

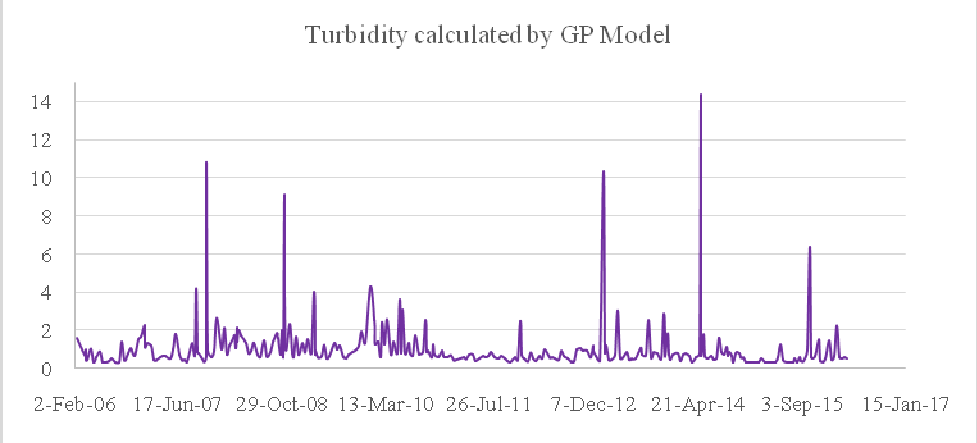

Figure 3: Turbidity of drinking water mesered at Klokot spring

In order to evaluate given GP model, it is also developed 5 different regression models with different kind of polynomial degree, and interactions of predictors. The following table shows calculated performance parameters of the GP model and 5 regression models.

Table 1: Comparison analysis of GP model and 5 diferent regression models

NR	Models	RMSE	RSE	SE	R	R^2
1	GP Model	0.868	14.732	217.022	0.723	0.850
2	RM_1	0.928	15.753	248.167	0.683	0.826
3	RM_2	1.020	17.303	299.401	0.617	0.785
4	RM_3	0.838	14.218	202.154	0.741	0.861
5	RM_4	0.800	13.580	184.419	0.764	0.874
6	RM_5	0.692	11.742	137.867	0.824	0.908

The following table shows performance criteria for testing data setwhich can show if the GP model is overfitted.

Table 2: Prediction analysis of GP model and 5 diferent regression models

NR	Models	RMSE	RSE	SE	R	R^2
1	**GP Model**	**0.395**	**0.967**	**0.935**	**0.059**	**0.243**
2	RM_1	0.377	0.923	0.852	0.113	0.336
3	RM_2	0.507	1.241	1.540	0.001	0.024
4	RM_3	0.456	1.117	1.248	0.001	-0.030
5	RM_4	0.637	1.561	2.436	0.033	-0.181
6	RM_5	0.636	1.558	2.426	0.052	-0.229

From the Table 2 we can see that RM_5, RM_4 and RM_3 have negative correlation which indicates no relation between calculated values and values from training data set. On the other hand, GP Model has the best value for most of the performance criteria.

4. SUMMARY

The paper shows development of genetic programming model for turbidity prediction in drinking water. Turbidity was measured from Klokotspring the main water supply source of Bihac town. Turbidity in water from Klokot spring is affected by many factors, mainly by the structure of the soil and groundwater around the spring. Turbidity measurements is collected in period from 2006 to 2016. Based on the collected data the GP model of turbidity is developed based on several predictors mentioned in the paper. For comparison analysis, several classic regression models for turbidity are developed with the same predictors. Since the turbidity values show nonlinear dependency and correlation to the predictors the 5 different regression models are developed in order to be compared with GP model. The analysis was shown that model calculated using GP method has better performance in comparison of models calculated by nonlinear regression analysis. As can be seen from the comparison analysis presented in tables 1 and 2, for nearly same performance criteria of the models, GP model gives better prediction and less overfitting. One of the main constrain for the regression models is predefined polynomial form. This means the polynomial form must be known before the modelling process started. This includes the number of regression coefficients, and the degree of the polynomial. On the other hand,GP models have no constrains in “shape”, as well as in degree. Moreover, GP Models can contains any function from the function set, as well as any combination of the functions and terminals, which leads the model can be highly nonlinear, and calculate better result than any regression models.

5. REFERENCES

[1] Mann, A. G., Tam, C. C., Higgins, C. D., & Rodrigues, L. C.:*The association between drinking water turbidity and gastrointestinal illness: a systematic review*, BMC Public Health, 7, 256, 2007, http://doi.org/10.1186/1471-2458-7-256.

[2] ISO 7027-1:2016*Water quality -- Determination of turbidity Part 1: Quantitative methods*, International Organization for Standardization, 2016.https://www.iso.org/standard/62801.html.

[3] BAS EN ISO 7027:2002Water quality - determination of turbidity, Institute for standardization ofBosnia and Herzegovina.

[4] ASTM D1889-00, *Standard Test Method for Turbidity of Water (Withdrawn 2007)*, ASTM International, West Conshohocken, PA, 2000, www.astm.org.

[5] Bonacci O.,*Hidrološkaanalizapojavemutnoćenaizvorima u kršu: Interpretacijapodatakamjerenihnaizvoru Omble*. Hrvatskevode 24, 47-57, 2016.

[6] Nebbache, S.,Feeny, V.; I. Poudevigne, I.; Alard, D.: *Turbidity and nitrate transfer in karstic aquifers in rural areas: The Brionne Basin case-study*. Journal of Environmental Management, 62, 389-398, 2001.

[7]Dedić S. idrugi:Predviđanjemutnoćevodeizvorišta Trbljevikmetodomgenetskogprogramirnaja,VMeđunarodnikongres “Inženjerstvo, ekologijaimaterijali u procesnojindustriji”, Jahorina 2017.

[8] KozaJ.R., *Genetic Programming: A paradigm for genetically breeding population for computer programs to solve problems*, Stanford : Computer Science Department, Stanford University, 1990.

[9] Hrnjica B, GPdotNET V4.0- artifical intelligence tool [kompjuterski program], http://github.com/bhrnjica/gpdotnet, zadnja posjeta 01/06/ 2017.

[10] Wolfram Research, Inc., Mathematica, Version 10.02, Champaign, IL (2014).

[11] Danandeh Mehr, A., Nourani, V. A Pareto-optimal moving average-multigene genetic programming model for rainfall-runoff modelling. Environmental Modelling & Software, 92, 239-251, 2017.