Ecological Engineering 100 (2017) 63–72

Contents lists available at ScienceDirect

Ecological Engineering

journal homepage: www.elsevier.com/locate/ecoleng

Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube

Anita Csábrági[a,*], Sándor Molnár[a], Péter Tanos[a], József Kovács[b]

[a] Institute of Mathematics and Informatics, Szent István University, Páter K. u.1, Gödöllő, H-2103, Hungary
[b] Department of Physical and Applied Geology, Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest, H-1117, Hungary

ARTICLE INFO

Article history:
Received 8 August 2016
Received in revised form 8 December 2016
Accepted 16 December 2016
Available online 24 December 2016

Keywords:
Dissolved oxygen forecasting
General regression neural networks
Multilayer perceptron neural networks
Multivariate linear regression
Radial basis function neural network

ABSTRACT

Dissolved oxygen content is one of the most important parameters in the characterization of surface water conditions. Our goal is to make a forecast of this parameter in Central Europe's most important river with the use of other, easily measurable water quality parameters (pH, temperature, electrical conductivity and runoff) with the use of linear and nonlinear models. We adapt four models for forecasting dissolved oxygen concentration, namely a Multivariate Linear Regression model, a Multilayer Perceptron Neural Network, a Radial Basis Function Neural Network and a General Regression Neural Network model. Data is available for Hungarian sampling locations on River Danube (Mohács, Fajsz and Győrzámoly) for the period of 1998–2003. The analysis was performed with four alternative combinations, the models were formulated using data from the period 1998–2002 and a dissolved oxygen forecast was made for 2003. Evaluating model performance with various statistical measures (root mean square error, mean absolute error, coefficient of determination, and Willmott's index of agreement), we found that non-linear models gave better results than linear models. In two cases the General Regression Neural Network provided the best performance, in two other cases the Radial Basis Function Neural Network gave the best results. A further goal was to conduct a sensitivity analysis in order to identify the parameter with the highest influence on the performance of the created models. Sensitivity analysis was performed for the combination of all three sampling locations (4th combination) and it was found that for all three neural network models sensitivity analyses showed that pH has the most important role in estimating dissolved oxygen content.

1. Introduction

To have a proper understanding of surface waters it is vital to know the water quality parameters provided by the data of monitoring networks. The operation of a monitoring network can be improved considering various criteria (e.g. cost efficiency), or can be facilitated by estimating certain parameters from other parameters. The current article examines several parameters which can be used for the estimation of dissolved oxygen.

Dissolved oxygen is a very significant parameter in the characterisation of the condition of aquatic ecosystems, thus forecasting its concentration using easily available and measureable parameters may be considered important scientific advantage. The sources of dissolved oxygen (DO) in a water body include re-aeration from the atmosphere, photosynthetic oxygen production, and DO loading. The sinks include the oxidation of carbonaceous and nitrogenous material, the oxygen demand of the sediment, and the respiration of aquatic plants (Kuo et al., 2007). The concentration of DO reflects the equilibrium, or the lack of one, between oxygen-producing and oxygen-consuming processes (Ahmed, 2014). Thus, the preservation of DO in water bodies is one of the primary concerns for water resource managers.

The estimation and forecasting of the major parameters of surface waters is typically performed using various types of artificial intelligence based techniques relying on machine learning. This requires training, validation (the latter can be omitted if data is

Abbreviations: ANN, Artificial Neural Network; C_A, C_B, C_C, C_D, Combination A, B, C, D; DO, dissolved oxygen; EC, electrical conductivity; GRNN, General Regression Neural Network; hydro PP, hydro power plant; HNPP, Hungarian Nuclear Power Plant; IA, Willmott's index of agreement; MAE, mean absolute error; MLPNN, Multilayer Perceptron Neural Network; MLR, Multivariate Linear Regression; R^2, coefficient of determination; RBFNN, Radial Basis Function Neural Network; RF, runoff; rkm, river kilometres; RMSE, root mean square error; WT, water temperature.

* Corresponding author.
E-mail addresses: csabragi.anita@gmail.com, csabragi.anita@gek.szie.hu (A. Csábrági), molnar.sandor@gek.szie.hu (S. Molnár), tanospeter@gmail.com (P. Tanos), kevesolt@gmail.com (J. Kovács).

http://dx.doi.org/10.1016/j.ecoleng.2016.12.027

scarce) and test sets. The creation of these sets can be undertaken various ways. Most mainstream sources suggest random creation of the respective sets (1), in this case the term 'estimation' or 'modelling' should be used (Ahmed, 2014; Antanasijevic et al., 2014; Basant et al., 2010; Emamgholizadeh et al., 2014; Heddam, 2014; Rankovic et al., 2012, 2010; Talib and Amat, 2012; Wen et al., 2013). Other sources divide sets according to sampling points (2), assigning the majority of sampling points to the training set, and a smaller proportion of sampling points to the test set; in this case the correct terminology is 'spatial forecasting' (Dogan et al., 2009; He et al., 2011a; Palani et al., 2008). Finally, some sources divide the temporal interval of measurement by assigning multiple initial years to the training set and a couple of final years to the test set (3), in this case 'temporal forecasting' is performed (Antanasijevic et al., 2013; Ay and Kisi, 2012; Csábrági et al., 2015; He et al., 2011b; Singh et al., 2009). This article proposes examples for the 3rd case, temporal forecasting.

In the following some results are presented from temporal forecasting studies. Antanasijevic et al. (2013) compared three Artificial Neural Networks (ANNs), namely, Multilayer Perceptron Neural Network (MLPNN), General Regression Neural Network (GRNN) and Recurrent Neural Network with Multivariate Linear Regression (MLR) to the forecasting of DO in the River Danube at a single location in Serbia, Bezdan. The data from the years 2004–2008 were used as a training data set, and the data from 2009 were applied as the test data set. The authors found that the Recurrent Neural Network performed much better than the others. Singh et al. (2009) developed two MLPNN models to forecast the biological oxygen demand and DO concentration in the River Gomti, India, with the help of 11 input water quality parameters. The entire water quality data set spanning 10 years was divided into three sub-sets; the training set contained data from the first 6 years, the validation set comprised data of the next 2 years, and the test data set consisted of the data from the remaining final 2 years. The authors established that the MLPNN was a powerful predictive tool in the computation of water quality parameters. Ay and Kisi (2012) developed and compared two ANNs – the MLPNN and RBFNN – and MLR for the forecasting of DO concentration by using four parameters (temperature (WT), pH, electrical conductivity (EC) and runoff (RF)) as input in Foundation Creek, Colorado, USA. The whole data set was collected from upstream and downstream USGS stations and the training, validation and test data sets were divided by date of the experimental data set. Comparison of the results showed that the Radial Basis Function Neural Network (RBFNN) model performed better than the MLPNN and MLR models, and that the RBFNN model was quite effective without the runoff parameter in DO concentration forecasting. Finally, the downstream DO concentration was successfully forecasted using only water temperature data of the upstream station. He et al. (2011b) applied MLPNN and MLR to forecast the daily DO minimum and the daily DO variation in the Bow River, Canada. The water quality parameters of 2006–2007 – recorded at 15 or 30 min intervals – of both sampling sites were used for the training set and the test set contained the data from 2008. The DO minimum was forecast using water temperature and runoff, and the input parameters for the estimation of daily DO variance were radiation, water temperature and runoff. In both cases the MLPNN model outperformed the linear model.

Our main goal is to aid water quality management using estimation procedures which optimise the operation of monitoring by ensuring cost efficiency and representativity. This may be attained by providing forecasts of DO-concentration, which is one of the most important hydrochemical parameters, using easily measureable physical and chemical parameters. We use the mainstream approaches to reach our objective, i.e. MLR, and the various ANN methods, and we provide (1) an efficiency ranking of these methods for different combinations of sampling locations (see the details in Section 2.1). (2) We examine if there is a difference in the efficiency of the respective estimations of the alternative combinations. For reasons of economy, the reduction of the number of parameters used could be considered; in this case, the models discussed can effectively support decision-making. (3) Sensitivity analysis was performed to identify those parameters with the highest impact on estimation for all three non-linear models and the results were evaluated.

2. Material and methods

2.1. Study area

The River Danube is a very important ecological and economic factor in the region. Thus, the conservation and improvement of its water quality is of primary importance to the future of the region. The Danube is the second longest river in Europe, with a length of 2817 km from the Black Forest (Germany) to the Black Sea (Romania). The section in Hungary is 417 km long, with an average RF of 2000 m^3/s. The construction of the Gabčikovo hydro power plant (hydro PP) on the Slovakian-Hungarian border significantly altered the Danube, with around 80% of the discharge being rerouted to the Slovakian side and a RF of only 400 m^3/s remaining on the Hungarian side. The river returns to its original riverbed at 1806 river kilometres (rkm) (Kovács et al., 2015b). An additional noteworthy anthropogenic impact of the Hungarian Nuclear Power Plant (HNPP) (Paks, at 1526 rkm) is the effect of the effluent coolant water reaching the Danube, thus the effluent RF of the HNPP has a direct effect on primary greenhouse gas emissions from the electricity grid (Molnár, 2002).

There are 12 sampling sites in the section of the River Danube flowing through Hungary (Fig. 1). The Mohács station (D11, 1451.7 rkm) was chosen as an "undisturbed" representative location, because this sampling site is not disturbed by tributaries or anthropogenic installations. The sampling location D11 belongs to the Section type 6 according to the results of Sommerhäuser et al. (2003) and Liška et al. (2015). Further two locations, Győrzámoly (D2, 1806.2 rkm) and Fajsz (D9, 1507.6 rkm) were considered "noisy" locations and classified to the Section type 4 and the Section type 5, respectively. The D2 sampling location is the first after the sub-channel of the Gabčikovo hydro PP rejoins the Danube, while D9 is the first sampling location after HNPP. The four combinations used for the analysis were as follows, the first three considered the individual data from Mohács, Fajsz and Győrzámoly, denoted respectively as C_A, C_B and C_C. Finally, the fourth combination simultaneously assessed the data from all three sampling locations (C_D).

2.2. Water quality data set

The complete river water quality data set was divided into two subsets. Data from 2003 were used as the test data set (26 data samples on all sampling locations), and data from 1998 to 2002 were used as the training set (128 data samples in D11, 125 data samples in D9 and 130 data samples in D2). The same training and test sets were used in the application of each model.

The models required input parameters (in our case, measured pH, RF (m^3 sec^{-1}), WT (°C), EC ($\mu S\,cm^{-1}$)) to generate the output (forecasted DO ($mg\,L^{-1}$)). Input and target data (measured DO) were entered into the applied models after the z-score normalization technique had been applied (normalizing so the inputs and targets have zero mean and unit standard deviation). The target parameter corresponding to the input parameters belonged to the same water sample, measured at the same time and location.

The descriptive statistics of the available data (Table 1) highlighted the fact that the parameters with the highest variance are