SZENT ISTVÁN
EGYETEM

SZENT ISTVÁN UNIVERSITY

Estimation of dissolved oxygen content of surface waterflows using neural networks

Theses of doctoral (PhD) dissertation

Anita Csábrági

Gödöllő
2019

Doctoral school **Denomination**:	Mechanical Engineering PhD School
Science:	Agricultural Engineering
Head of school:	Prof. Dr. Farkas István professor, DSc Szent István University, Faculty of Mechanical Engineering, Institute of Environmental Systems
Supervisor:	Prof. Dr. Molnár Sándor professor, DSc Szent István University, Faculty of Mechanical Engineering, Institute of Mechanical Engineering Technology
Co-supervisor:	Dr. habil. Kovács József associate professor, PhD Eötvös Loránd University, Faculty of Sciences, Department of General and Applied Geosciences

....	....
affirmation of head of school	affirmation of supervisors

TABLE OF CONTENTS

NOMENCLATURE

Symbol	**Description**	**Unit**
CCDA	combined cluster and discriminant analysis	[-]
DO	dissolved oxygen	[mg L^{-1}]
EC	electrical conductivity	[μScm^{-1}]
GRNN	general regression neural network	[-]
IA	index of agreement	[-]
MAE	mean absolute error	[-]
MLPNN	multilayer perceptron neural network	[-]
MLR	multivariate linear regression	[-]
MSE	mean square error	[-]
Q	runoff	[m^{3}s^{-1}]
R^2	coefficient of determination	[-]
RBFNN	radial basis function neural network	[-]
RMSE	root mean square error	[-]
T_w	water temperature	[^{o}C]

1. INTRODUCTION AND OBJECTIVES

In this chapter the significance of my topic and the objectives of my research are presented.

1.1. Timeliness and relevance of the topic

In the last century the intensification of industrial and agricultural activity and urbanisation caused significant pollution in our natural aquifers and surface waters. This adverse anthropogenic impact caused the physical parameters (turbidity, light penetration, flow speed, temperature, etc.) reflecting the ecological condition of our waters to change drastically. To monitor and control the anthropogenic impacts it is vital and inevitable to have a precise knowledge of the water quality parameters provided by the river monitoring network. The operation of the monitoring network can be assessed or improved from multiple aspects (e.g. cost efficiency) if certain difficultly or expensively measurable variables are estimated from easily measurable variables. An adequate tool for this purpose is artificial intelligence or artificial neural networks models.

The last decades brought an increasing rate of application of artificial neural networks (ANNs) for the estimation of water quality variables either in lakes or rivers due to the inherent advantages they represent. For the application of neural networks no other necessity but input data is required. These models are also able to map the complex relationships between input and output data and generalise the experiences gained in an appropriate manner.

In most cases neural networks are used to examine the chemical parameters characterising natural water oxygen levels. Most frequently dissolved oxygen content (DO) is estimated as this is the most important parameter in natural waters and a vital indicator of surface waters' ecological balance.

The increased utilisation of neural networks for the estimation of dissolved oxygen motivated me to adapt this methodology for the two major domestic rivers while maintaining applicability to other rivers in terms results and conclusions.

1.2. Objectives

The objective of my dissertation is to present more efficient estimations by applying neural networks providing relevant cases of models and their configurations to forecast dissolved oxygen content of surface waters. In order to achieve this, I used sample data from the two major domestic rivers since to my latest knowledge estimation of DO content for these rivers was never implemented. Throughout my research I utilised three models of artificial neural networks and a multivariate linear regression model.

The one of the main aims of the research was for both rivers to explore which model provides efficient estimations for dissolved oxygen content.

My goal was also to provide temporal forecasts in order to assess the improvement of neural networks over a simple linear model.

Moreover I wanted to examine if the estimation is influenced by the impacts on the location of the respective measurements.

For the optimal spatial forecast of dissolved oxygen content my objectives were on the one hand to characterise geographically different sections of the river Tisza according to the efficiency of estimations of the respective models. On the other hand using the homogenous groupings of both rivers I wanted to assess which data structure would provide the highest efficiency for forecasting.

2. MATERIAL AND METHOD

In this chapter I present the results determining the homogenous groups of the examined river sections. Furthermore I discuss the models and their configurations and the chain of logic applied in my analysis which was used to implement my research goals.

2.1. Determination of homogenous groups on both rivers

My co-supervisor and his research team has undertaken a combined cluster and discriminant analysis (CCDA) to cluster the 12 sampling locations on the River Danube into homogenous groups. This resulted in 7 single element groups (D1, D2, D3, D4, D5, D8 and D9), one bi-element group (consisting of D6 and D7) and one triple-element group (which includes D10, D11 and D12) (Fig. 1).

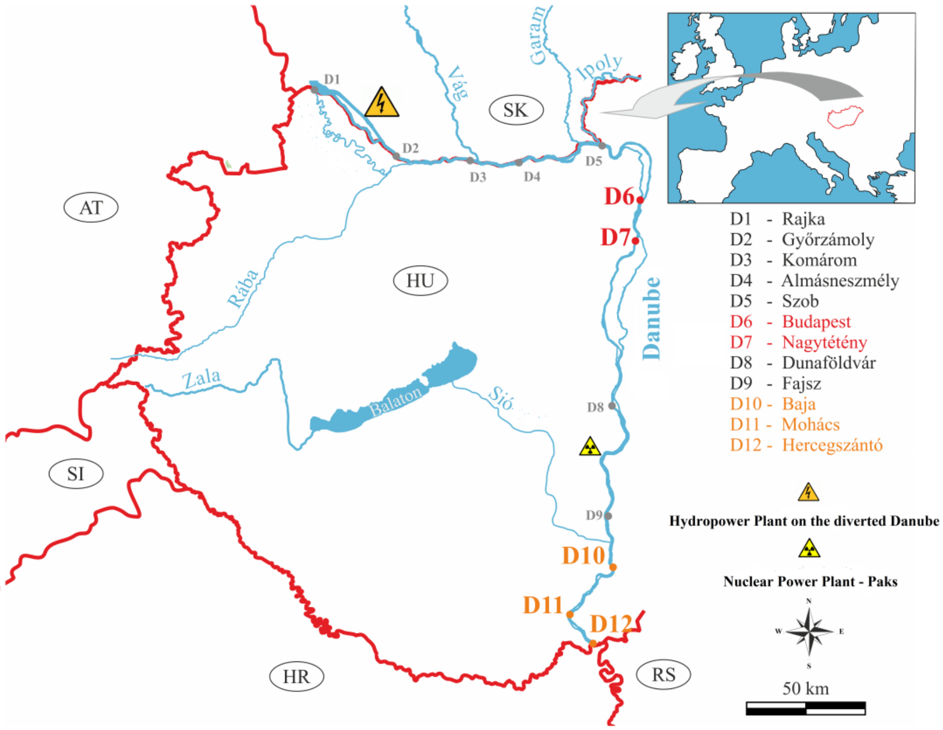


Fig. 1. Homogenous groups defined by CCDA on the River Danube

The homogenised groups created by CCDA on the River Tisza from my co-supervisor's research was utilised in my research. This result classified the 13 sampling locations into 10 homogenous groups. Three groups contained two locations, the remaining groups were of a single element

(Fig. 2). Namely, the upper section's T03-T04 locations, the mid-section's T07-T08 and the lower section's T11-T12 sampling locations formed the two-element homogenous groups. Stations with identical colours belong to the same homogenous groups.

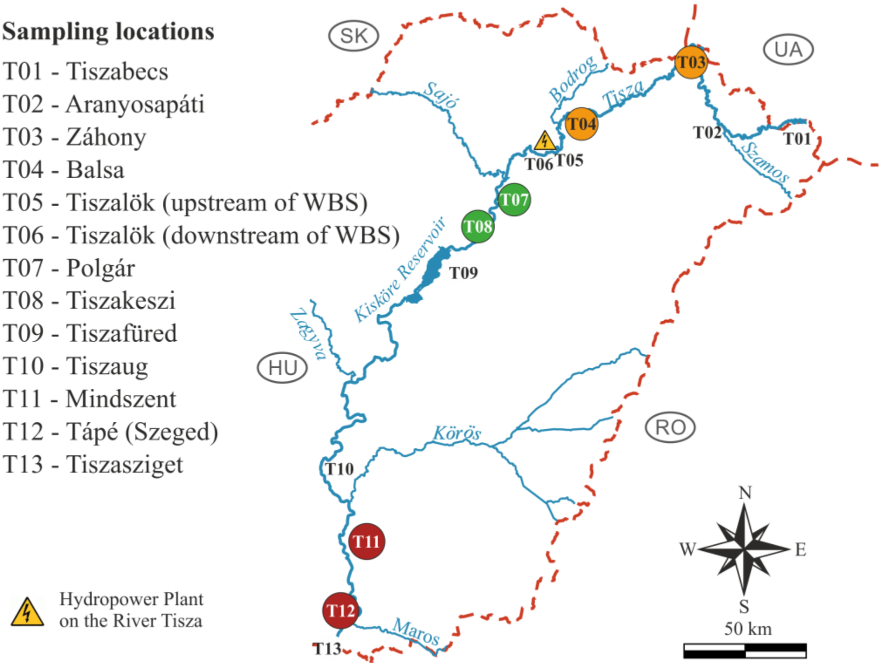


Fig. 2. Homogenous groups defined by CCDA on the River Tisza

2.2. Logics of assessment

In my analysis I estimated the dissolved oxygen content of the respective rivers from four parameters: water yield (Q, m^3s^{-1}), temperature (T_{w}, ^{o}C), water pH, and electric conductivity (EC, µScm^{-1}). The time interval was the period between 1998 and 2003.

The following three major steps were the cornerstones of my assessment:

1. Temporal forecast for dissolved oxygen content of the River Danube
   a) According to my objectives I examined how the presence of anthropogenic impacts on the sampling locations influence the estimation. This was possible since two power plants are located on the Hungarian section of the river (Fig. 1). For this purpose I have created four combinations: I estimated the DO-concentration separately at the stations at Mohács (C_A), Fajsz (C_B) and Győrzámoly (C_C) then for the composition of all these stations (C_D).

Four models were applied: multivariate linear regression (MLR), multilayer perceptron (MLPNN), radial basis function neural network (RBFNN) and general regression neural network (GRNN). For all versions, the data for the test set came from the year 2003.

2. Spatial forecast of the DO-concentration on the River Tisza.

   a) The examinations for the River Tisza were performed in three configurations, the first was basically a reference model (TC1) used to benchmark the results of other configurations. In the TC1 model the total dataset from the river was randomly distributed between the training and test sets, so this case actually was a model for the total domestic section's DO-concentration.

   b) The heterogeneous nature of the river's domestic section (significant differences can be observed in input parameters and dissolved oxygen levels) justified the examination and characterisation of the distinct river sections, this was the reason for the creation of the second configuration (TC2). In this configuration (TC2) 4 adjacent locations out of the total 13 sampling locations were used for the test set while the remaining 9 locations were used for the training set. In this case a spatial forecast was undertaken for the respective river section's DO-concentration. I used four alternative setups for the configuration (TC2-A, TC2-B, TC2-C and TC2-D) as follows. I tested first with the upper four sampling stations on the river (T01, T02, T03 and T04, this was denoted as TC2-A). Then, keeping the last station I added the next three stations to create TC2-B (T04, T05, T06 and T07). I continued in this manner, retaining only the last station in the set and adding the next three stations to formulate TC2-C (consisting of T07, T08, T09 and T10) and TC2-D (consisting of T10, T11, T12 and T13) respectively.

   c) Besides spatial characterisation I wanted to improve the efficiency of the estimation by assessing the data structure using the results of the CCDA method on the River Tisza. This formulated the basis of the third configuration (TC3) which had three setups (TC3-A, TC3-B and TC3-C) as the river has three bi-element homogenous groups (Fig. 2). For all setups I considered the two homogenous sampling points and their neighbouring inhomogeneous sampling location downstream. So from each of the upper, middle and lower sections of the river three stations were selected and examined (T03, T04 and T05, as setup „A"; T07, T08 and T09, as setup „B", and T11, T12 and T13, as setup „C”). In all cases two arbitrary stations were used for training set and one for test set. This meant three additional cases for the three setups altogether resulting in nine cases. In this configuration spatial optimisation was also implemented besides spatial forecasting.

d) For all three configurations on Tisza I applied three models (MLR, RBFNN and GRNN) and strived to maintain the 2:1 proportion for the training and test sets in all configurations.

3. Spatial forecast of DO-concentration based on data from the River Danube

a) In this case a reference model was created (DR) where the total data set was distributed between the training and test set in a 2:1 proportion, thus DO-concentration was modelled for the whole Hungarian section of the river.

b) Besides the reference model I developed a configuration where three stations were examined (D6, D7 and D8 – Budapest, Nagytétény and Dunaföldvár respectively) of which two stations were homogenous and the third was the neighbouring inhomogeneous stations (Fig. 1). In all cases two of the three stations were in the test set and one in the training set. First D6 and D8 (case DA) then D7 and D8 (case DB) finally D6 and D7 (case DC) were the elements of the training set. The objective was to determine the most efficient data structure of the training set.

c) Three models were applied to all configurations: MLR, RBFNN and GRNN.

3. RESULTS

This chapter presents the estimations, evaluations and comparison of modelling results from linear and neural network models on the DO-concentration of the Rivers Danube and Tisza.

3.1. Temporal forecast on the River Danube

3.1.1. Modelling results of the examined combinations

The modelling results (Table 1) show that of all four combinations (see section 2.2) the best performer was GRNN model in combination C_A. Differences between models created for the respective combinations show that RBFNN model provided the smallest RMSE values for C_B and C_D (Table 1). For combination C_C the smallest test set RMSE value was provided by the GRNN model. For all configurations it is visible that DO-content can be estimated more efficiently with neural networks than with a linear model. The two most efficient models from the three neural network models were GRNN and RBFNN. They both gave better results than MLPNN which underperformed in almost all cases. Another advantage of the aforementioned two models was their much shorter runtime compared to the double iteration training method used by the MLPNN model.

Table 1. Model performance on the test set in the examined combinations

Combination	Model	MLR	MLPNN	RBFNN	GRNN
C_A	RMSE [mg L^{-1}]	2.03	1.57	1.65	1.42
	R^2	0.4	0.57	0.59	0.72
C_B	RMSE [mg L^{-1}]	1.94	1.72	1.62	1.74
	R^2	0.44	0.58	0.54	0.55
C_C	RMSE [mg L^{-1}]	1.57	1.46	1.43	1.36
	R^2	0.5	0.59	0.47	0.6
C_D	RMSE [mg L^{-1}]	1.98	1.7	1.63	1.7
	R^2	0.41	0.59	0.59	0.55

3.1.2. Influence of anthropogenic effects on estimation

If the RMSE and R^2 indicators of neural networks are compared with MLR in all four configurations (Table 2.) it is visible that the largest improvement – 30% - is achieved by applying neural networks in configuration C_A in the undisturbed sampling location (Mohács). Neural networks are also significantly more efficient in other configurations but this scale of improvement could only be observed on undisturbed sampling locations. For noisy sampling points the anthropogenic impacts make estimation more difficult. In the case of sampling location D2 the neighbouring small hydro plant significantly influences water run-off. In the case of sampling location D9 the proximity of a nuclear power plant has a large influence on water temperature and thus on DO-values. These factors make DO forecasting more difficult in D2 and D9 locations. At the same time for the C_D combination a complex system was modelled which contained undisturbed and noisy sampling locations simultaneously. This explains why in this case DO forecasting was more difficult compared with C_A. The improvement of RMSE values in configuration C_D over the MLR values is only 14% as compared with the 30% improvement observed in configuration C_A (Table 2).

Table 2. RMSE and R^2 values in percent of the values of the MLR model

RMSE	MLR	MLPNN	RBFNN	GRNN	$\mathbf{R^2}$	MLR	MLPNN	RBFNN	GRNN
C_A	100%	77%	81%	70%	C_A	100%	143%	148%	180%
C_B	100%	89%	84%	90%	C_B	100%	133%	123%	126%
C_C	100%	93%	91%	87%	C_C	100%	117%	94%	119%
C_D	100%	86%	82%	86%	C_D	100%	135%	143%	143%

Results of the analysis have shown that all three neural networks and especially GRNN and RBFNN are efficient tools in the estimation of surface water dissolved oxygen concentration even if anthropogenic impacts influence these rivers. In this latter case the potential improvement of estimation efficiency is smaller.

3.2. Spatial forecast on the RiverTisza

3.2.1. Model results of the examined configurations

Along the analysis of the River Tisza the training and test sets in configuration TC1 were allocated randomly from the total dataset in a 2:1 proportion. Comparing results from the three randomly (initialised with 800, 2000 and 1000 initial values) generated selections (TC1-A, TC1-B and TC1-C) - as shown in Table 3 - it can be seen that the best results came from the GRNN model in TC1-B combination. Therefore this will be our reference model and its RMSE value of 2.14 mg L^{-1} will be the reference value (Table Table 3.) used to benchmark the other configurations' results with.

Table 3. Results of configuration TC1 for the test set

Setup	Model	MLR	GRNN	RBFNN
TC1-A	RMSE [mg L^{-1}]	2.40	2.33	2.38
	R^2	0.35	0.39	0.39
TC1-B	RMSE [mg L^{-1}]	2.25	2.14	2.23
	R^2	0.41	0.46	0.41
TC1-C	RMSE [mg L^{-1}]	2.35	2.31	2.26
	R^2	0.40	0.42	0.44

In configuration TC1, the total domestic section of the river is estimated in a single step with the created models. This assumes that the total river section has uniform attributes so a single model can describe the DO-concentration of the river. This assumption however does not hold for lengthy sections of the large rivers. This necessitated a geographically controlled distribution of data into training and test sets leading to the introduction of a second and a third configuration.

In configuration TC2, I also used the total dataset for the assessment. I used 9 sampling locations for training and four neighbouring locations for testing in order to develop separate models for the distinctive river sections and I applied four setups (section 2.2.).

The worst performance was observed in setup TC2-A with all three models (Table 4). The best result (1.64 mg L^{-1} RMSE) was provided by the GRNN model in setup TC2-C. For setups TC2-A and TC2-B the GRNN model gave the best results in terms of RMSE, while for setup TC2-D RBFNN model showed the best performance.

The setups for configuration TC2 gave a more accurate estimation for certain sections than the reference model. The modelling results from the upper section showed worse performance (setups TC2-A and TC2-B) than the reference model. The forecast of the lower sections in setups TC2-C and TC2-D showed however a higher efficiency than the reference model (summary chart, Fig. 3). The size of the output domain of the test set in TC2-A was 1.9-28.5 mg L^{-1}, while the size of the training set output domain was 3.6-14.2 mg L^{-1}. Since neural networks cannot give a proper estimation for values out of range therefore the unfavourable results for setup TC2-A are not surprising.

Table 4. Results of configuration TC2 on the test set

Setup	Model	MLR	GRNN	RBFNN
TC2-A	RMSE [mg L^{-1}]	4.75	4.73	4.74
	R^2	0.07	0.06	0.08
TC2-B	RMSE [mg L^{-1}]	2.66	2.56	2.59
	R^2	0.20	0.25	0.26
TC2-C	RMSE [mg L^{-1}]	1.65	1.64	1.75
	R^2	0.67	0.66	0.50
TC2-D	RMSE [mg L^{-1}]	1.71	1.81	1.69
	R^2	0.70	0.61	0.60

Configuration TC2 was necessary for the spatial characterisation of the river nevertheless its implementation was not efficient enough for the upper sections. In order to improve estimation efficiency I applied spatial optimisation as results described in section 2.1 were available which were used in configuration TC3 according section 2.2.

In configuration TC3, I did not use the total dataset of the river but worked only with three stations' data at a time with two stations being of a homogenous group and their downstream neighbour as the third station. The training set elements were composed of data from two sampling points and the third station's data was used for the test set.

I examined three cases (section 2.2): three sampling locations from the upper section of the river (T03, T04, T05, this was setup „A”, Table Table 5.), here stations T03 and T04 formed the homogenous group while T05 had a different structure.

Table 5. Results of configuration TC3 setup „A" for the test set

Sub-setup	Training station + test station	Model	MLR	GRNN	RBFNN
TC3-A#1	T3, T4 + T5	RMSE [mg L^{-1}]	3.93	3.70	3.58
		R^2	0.65	0.42	0.34
TC3-A#2	T3, T5 +T4	RMSE [mg L^{-1}]	3.49	3.12	3.05
		R^2	0.06	0.22	0.24
TC3-A#3	T4, T5 + T3	RMSE [mg L^{-1}]	3.35	2.93	2.97
		R^2	0.06	0.24	0.22

At the mid-section of the river I dealt with one setup (T07, T08, T09, denoted as setup „B" Table 6) where T07 and T08 were homogenous while sampling location T09 had a different structure.

Table 6. Results of configuration TC3 setup „B" for the test set

Sub-setup	Training station + test station	Model	MLR	GRNN	RBFNN
TC3-B#1	T7, T8 + T9	RMSE [mg L^{-1}]	1.25	1.17	1.23
		R^2	0.76	0.78	0.74
TC3-B#2	T7, T9 +T8	RMSE [mg L^{-1}]	1.01	0.84	0.90
		R^2	0.78	0.84	0.83
TC3-B#3	T8, T9 + T7	RMSE [mg L^{-1}]	1.00	0.84	0.89
		R^2	0.77	0.84	0.82

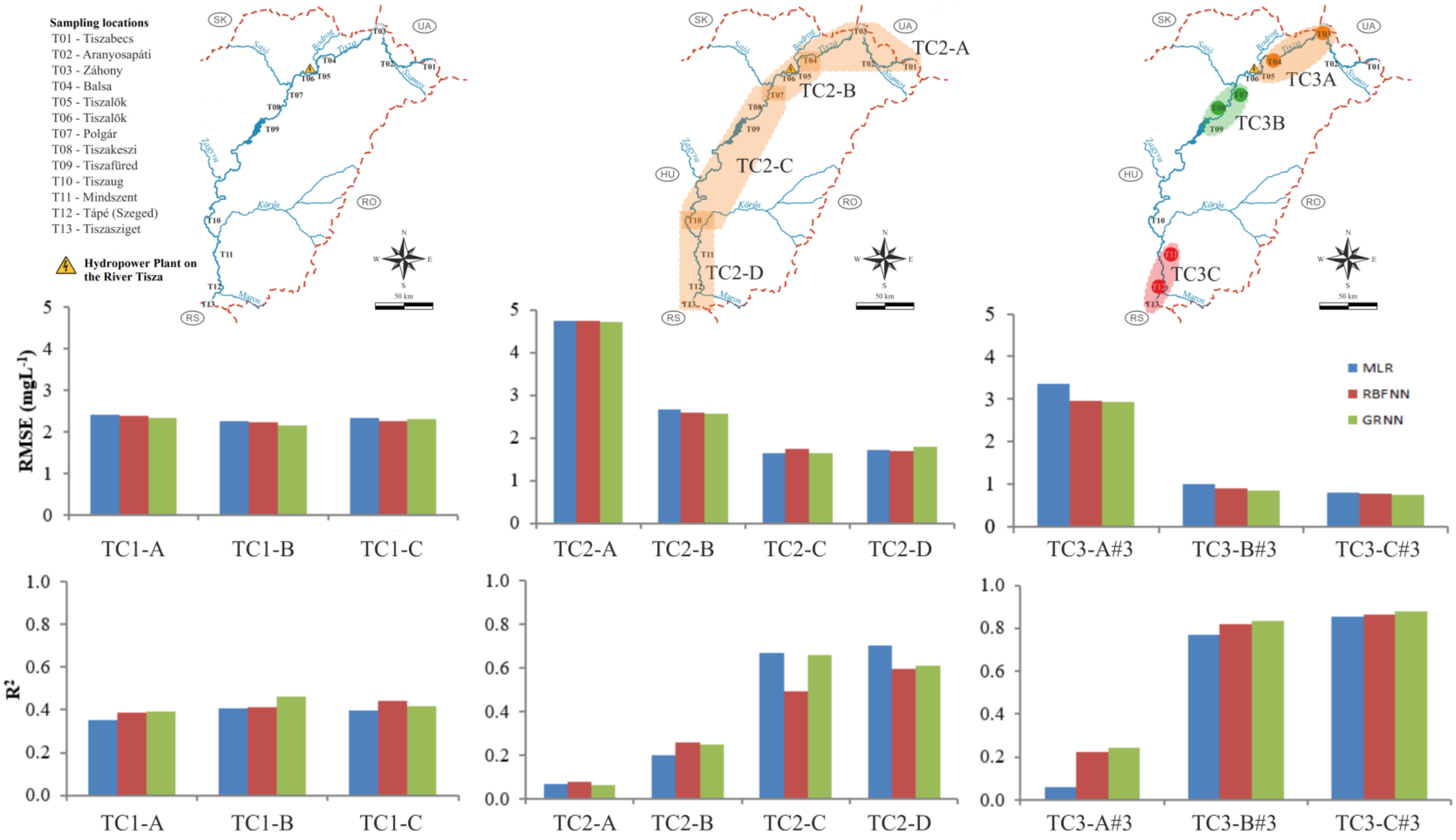
Finally, I examined the last three domestic stations of the river (T11, T12 and T13, this is setup „C", Table 7) where T13 had a different structure from the homogenous group formed by stations T11 and T12.

Table 7. Results of configuration TC3 setup „C” for the test set

Sub-setup	Training station + test station	Model	MLR	GRNN	RBFNN
TC3-C#1	T11, T12 + T13	RMSE [mg L^{-1}]	0.87	0.77	0.86
		R^2	0.83	0.86	0.82
TC3-C#2	T11, T13 +T12	RMSE [mg L^{-1}]	0.95	0.75	0.76
		R^2	0.81	0.89	0.88
TC3-C#3	T12, T13 + T11	RMSE [mg L^{-1}]	0.80	0.74	0.77
		R^2	0.85	0.88	0.86

3.2.2. Selecting the most efficient model

The built-in maps and results of all three configurations are presented on a summarising figure (Fig. 3) which shows the RMSE and R^2 values of the test set. For the TC3 configuration as for all three setups the best results were achieved by the third sub-setups TC3-A#3, TC3-B#3 and TC3-C#3 (Table 5), Fig. 3 shows these values. For the sub-setups TC3-A#3, TC3-B#3 and TC3-C#3 the GRNN model gave the most efficient estimations in terms of both statistical indicators.



a) configuration TC1

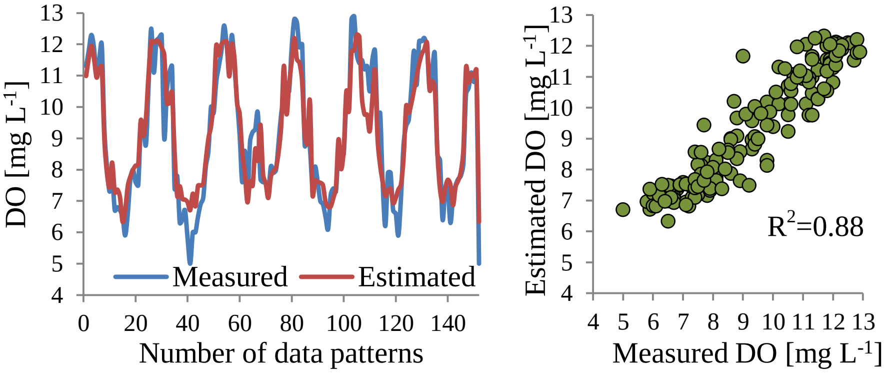
b) configuration TC2

c) configuration TC3

Fig. 3. Installed maps and results of Tisza configurations

In configuration TC3, I examined three different setups (TC3-A, TC3-B and TC3-C) with 3 separate sub-setups in each. RMSE values from TC3-A were less than 4 mg L^{-1} which were better than TC2-A results but still lagging behind reference values with 37%. This can be explained with the strong distinctive characteristics of T05 station due to the anthropogenic effect of the Tiszalök dam and hydro. The results from the other two setups (TC3-B and TC3-C) outperform reference values significantly thus estimation for the lower section is more efficient not only in the TC2 but in the TC3 configuration (Fig. 3).

Examining the results of the three configurations the ultimate best results were achieved in configuration TC3 with setup C (Table 7) with the GRNN model in the sub-setup TC3-C#3 where the calculated RMSE value was 0.74 mg L^{-1}, which was a 65% improvement over the reference value. Here data from station T11 (Mindszent) formed the test set (Fig. 4).



a) line chart b) scatterplot

Fig. 4. Diagrams of mesaured and estimated DO-levels at TC3-C#3

Assessment of the efficiency of the applied models by calculating an average RMSE and R^2 indicator for all three configurations shows that neural networks are more efficient means for estimation than multivariate linear regression (Fig. 5). The GRNN model turned out to be the most efficient based on the results of configurations on the River Tisza.

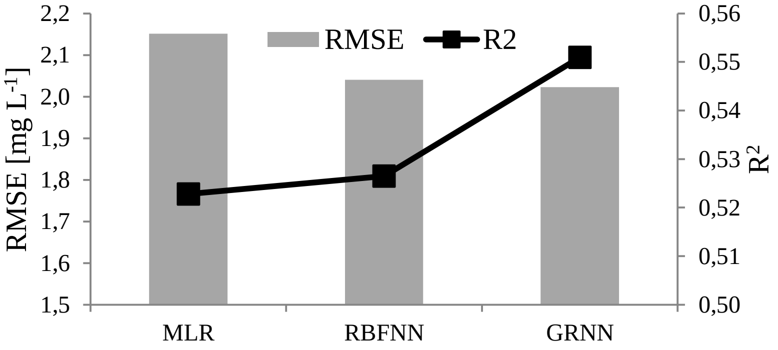


Fig. 5. Average values of RMSE and R^2 values from the assessed configurations

3.2.3. Efficient data structure of the training set on Tisza

Results from setups A, B and C of configuration TC3 underlined that the worst results come from those sub-setups where only the homogenous group's sampling points are included in the training set while the inhomogeneous location is put in the test set (Table Table 5. , TC3-A#1, TC3-B#1, TC3-C#1). Therefore model efficiency can be improved by including both a homogenous and inhomogeneous station in the training set so the latter contains both data structures making estimation more efficient. In this case the training set is of „mixed structure", in the followings this is the notation I will use.

Results have shown that despite the reduction in sample size the conscious assignment of training and test sets and spatial optimisation can still improve model efficiency significantly.

3.2.4. Characterisation of spatial sections of the river

Comparing the results of configurations TC2 and TC3 with the reference model it was visible that the upper section of the river gave ground to worse performing models (TC2-A, TC2-B and TC2-C setups) and this shows that the larger variability of the upper section makes estimation more difficult.

At the same time however, the efficiency of TC2-C, TC2-D and TC3-B, TC3-C which forecasted the lower sections of the river surpassed the values of the reference model (Fig. 3.).

Altogether the results of configurations TC2 and TC3 confirmed that it is easier to estimate DO concentration on the lower sections of the river. This can be traced back to the fact that with reduction of the river flow speed the variability of water quality parameters decrease making the estimation of the aquatic system easier. As a result the most accurate estimation was thus achieved on the lower section of the river (TC3-C#3, Fig. 4) applying a mixed structure.

3.3. Spatial forecast on the River Danube

3.3.1. Modelling results of the analysed configurations

I estimated the DO-concentration of Danube with four input parameters and in two alternative configurations with alternative methods of training and test set formation. In the first configuration I randomly selected data from the total sample set into the training and test sets maintaining a 2:1 proportion. In this case the DO-concentration forecast was made for the complete in-country section of the river. Results from the random selection (initial value of 2000) denoted as setup D_R are shown in Table 8.

Table 8. Results of Danube configurations for the test set

Setups	Training set+ Test set	Model	MLR	GRNN	RBFNN
D_R	2/3 + 1/3	RMSE [mg L^{-1}]	1.39	1.22	1.27
		R^2	0.33	0.50	0.45
D_A	D6, D8 + D7	RMSE [mg L^{-1}]	1.27	0.93	1.08
		R^2	0.48	0.69	0.59
D_B	D7, D8 + D6	RMSE [mg L^{-1}]	1.41	0.96	1.01
		R^2	0.37	0.66	0.65
D_C	D6, D7 + D8	RMSE [mg L^{-1}]	1.89	1.79	1.75
		R^2	0.32	0.39	0.41

The results once again highlighted that neural networks provide more accurate estimation than multivariate linear models. The highest efficiency was achieved by GRNN models therefore the RMSE value for the test set from the GRNN model runs will be the reference value (1.22 mg L^{-1}).

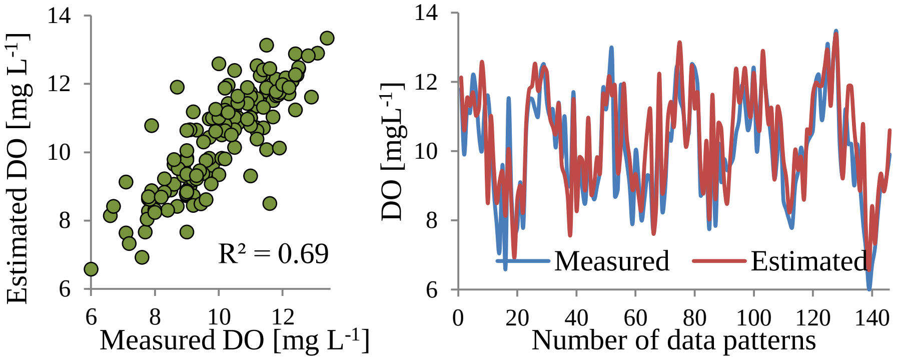
In the second configuration I could only use a single bi-element group from the homogenous groups on the Danube identified by the CCDA method (Fig. 1) to define the most efficient data structure of the training set. For this I needed two homogenous stations and their downstream neighbour inhomogeneous station. This was composed from the D6

(Budapest) and D7 (Nagytétény) homogenous stations and their downstream neighbour D8 (Dunaföldvár) which can be considered inhomogeneous (see section 2.1). Therefore one setup was examined with data from D6, D7 and D8 where the first two stations were the training set and the last was the test set thus maintaining the 2:1 proportion for the training and test set. Thus three sub-setups had to be examined namely where respectively D6 and D8 (sub-setup D_A), D7 and D8 (sub-setup D_B) and finally D6 and D7 stations (sub-setup D_C) were the elements of the training set.

For all three sub-setups I examined which model provided the smallest RMSE values for the test set, in other words which model was the most efficient. Accordingly the D_A setup and GRNN model gave the smallest RMSE value of 0.93 mg L^{-1} on the test set. For the D_B setup also the GRNN model was the most efficient (0.96 mg L^{-1}) while for the D_C setup the RBFNN model gave the lowest RMSE value (1.75 mg L^{-1}). Therefore in all three setups the two neural networks could achieve higher efficiency than the MLR model.

3.3.2. The most efficient data structure of the training set

Based on the result of the second configuration it can be seen (Table 8) that the worst result came from sub-setup D_C and although the RBFNN model was the most efficient here the RMSE values on the test set were respectively 88% and 82% larger (worse) than the best RMSE-values of setups D_A and D_B. According to section 3.2.3 the training set of these setups were of „mixed structure” as it consisted of data from a homogenous and an inhomogeneous station. Estimations from setups D_A and D_B show close results based on test set statistical indicators but overall the D_A setup and GRNN model provided the highest efficiency (Fig. 6) with Nagytétény station data used for test set.



a) scatterplot b) line chart

Fig. 6. Diagrams of measured and estimated DO-concentration at D_A

Best RMSE values for D_A and D_B setups showed a 31% and 27% efficiency improvement over the reference values nonetheless for the D_C sub-setup a 43% decline in efficiency could be observed compared to the reference model.

Since sample size in the second configuration is less than a quarter of the first configuration's size therefore the efficiency improvement of D_A and D_B sub-setups is obvious since the significant reduction of sample size (using spatial homogenity) still led to more efficient estimations. In the same time it should be noted that for spatial optimisation it matters how homogenous and inhomogeneous sampling locations are grouped into the training and test sets as a less efficient outcome can also be the result.

4. NEW SCIENTIFIC RESULTS

My research focused on the estimation of dissolved oxygen content of surface waters through the application of several artificial neural network methods. The concluding results can be summarised in the following theses:

1. Influence of anthropogenic impacts on estimation

Using data of the River Danube I made a temporal forecast for the dissolved oxygen content of the river using MLR, MLPNN, GRNN and RBFNN models. My results confirmed that the MLR-based reference model's estimation can be significantly (30%) improved on the undisturbed sampling location (Mohács) if neural networks are applied. In the case of the two other noisy sampling locations (Fajsz, Győrzámoly) a less impressive (13-18%) improvement can be achieved. Efficiency of estimation was measured by RMSE calculated on the test set. I ascertained that it is more difficult to provide a comparatively more efficient estimation with neural networks when data from sampling locations under anthropogenic effects are used.

2. Efficient data structure of neural networks' training set

Using data from Tisza and Danube I undertook a spatial forecast and spatial optimisation with respect to dissolved oxygen content using MLR, GRNN and RBFNN models considering homogenous groups for training and test sets. This meant three cases for Tisza and one case for the Danube. I used these results to demonstrate the higher efficiency of estimation based on the RMSE value of the test set when it is of „mixed structure", in other words, when it contains one of the elements of the bi-element homogenous group and an inhomogeneous location. In all four examples the inhomogeneous station was the downstream neighbour of the bi-element homogenous group.

3. Capability to provide spatial estimation for sections of the River Tisza

With the spatial optimisation and forecast of dissolved oxygen content of the River Tisza I have confirmed that on the Hungarian sections of the river this parameter can be estimated with high precision. On the upper, more turbulent and faster flowing section the estimations are less accurate (e.g. Balsa, Záhony, sections above the Tiszalök-dam) than on the lower sections (Mindszent, Tápé, Tiszasziget). This highlighted that the (upper, middle, lower) sectional characteristics of the rivers under examination should be considered for the estimations.

4. *Selecting the efficient model for the estimation of dissolved oxygen content*

My analyses have confirmed that neural networks – especially GRNN and RBFNN – are more efficient for dissolved oxygen content estimation in terms of test set RMSE values than multivariate linear regression. The performance gap between neural networks and multivariate linear models was larger for the examinations on Danube.

5. CONCLUSIONS AND SUGGESTIONS

From the results of the temporal forecasts on the Danube I concluded that it is very important to analyse the effects influencing the river on the given sampling location. Namely it seems that „noisy" stations influenced by anthropogenic effects are more difficult to estimate both with linear and neural network models while undisturbed stations data can provide a more precise and reliable estimation.

After studying publications on estimation of riverine waters' dissolved oxygen content I found that there are almost no scientific studies where the river sampling locations were considered during the formulation of training and test sets, in other words spatial forecast would be undertaken. In my research I give examples for spatial forecast on both the Danube and Tisza rivers.

During the analysis of the River Tisza I modelled with data from all the sampling points in the first two configurations and I concluded that it would be useful to examine the structure of the river's sampling locations in the third configuration. In this case I achieved more efficient estimations with the lower sample size optimised input dataset compared to the first two configurations. Therefore despite the reduced sample size the controlled assignment of training and test sets led to a significant improvement in model efficiency which does not contradict statistical consistency as the Hungarian section of the River Tisza cannot be considered structurally uniform.

The most efficient estimation was achieved using spatial optimisation and applying a mixed structure in the training set using one of the two homogenous stations and its neighbouring inhomogeneous station for its elements. The most efficient data structure for the training set is presented in three examples on the River Tisza and one example on the River Danube. In all four cases the two element homogenous group was coupled with the downstream neighbour inhomogeneous station. In my future research it would be reasonable to examine if the application of a mixed structure would also result in a more efficient outcome if the inhomogeneous station is not the neighbour of the - two element or larger - homogenous group.

As a continuation of the research it might be practical to use additional neural networks e.g. support vector machines, hybrid models, adaptive neuro-fuzzy models to estimate the dissolved oxygen content of the rivers.

6. SUMMARY

Due to the increasing anthropogenic industrial activity significant changes have accrued in natural water quality therefore the estimation of parameters best describing dissolved oxygen concentration of surface waters is an important task. My research aimed to present a more efficient application of neural networks for the estimation of this parameter on data of our two major domestic rivers, Danube and Tisza.

Throughout the literature review I presented the types and respective advantages and disadvantages of neural networks. I also identified those easily measurable parameters which influence dissolved oxygen content the most.

From the sampling locations I selected three stations according to the occurring anthropogenic effects. Mohács was the „unbiased" reference station while Győrzámor and Fajsz were the two „noisy" stations. The three stations were examined separately and in composition to implement a temporal forecast using results from four models (MLPNN, GRNN, RBFNN and MLR). Significant improvement was achieved by the GRNN model over MLR for data from Mohács. It turned out to be difficult for the two other noisy stations to provide a more efficient estimation over the linear model.

For all locations on the Tisza the sampled data were allocated into training and test sets with three methods. The first configuration was a reference model with random selection, the second provided a spatial forecast with selection based on the geographical location of sampling points. In the third configuration (spatial optimisation) only three stations' data were examined simultaneously. In all cases data were allocated in a 2:1 proportion between the training and test sets for comparative purposes and three models (GRNN, RBFNN and MLR) were applied. The results show that the upper section of the River Tisza is more difficult to estimate and the efficiency of the models improve downstream. I proved with results from the spatial optimisation setups on Tisza and Danube that if the training set is of mixed structure (consisting of homogenous and inhomogeneous sources) then the estimation is more efficient. A spatial optimisation resulted in 65% improvement on the lower section of Tisza compared to the reference model using a „mixed" structure training set. I found that neural networks – especially GRNN and RBFNN models – are efficient means to estimate dissolved oxygen content of the rivers than linear models.

7. PRIORITY PUBLICATIONS RELATED TO THE THESES

Proofread articles in an international language

1. **Csábrági A.,** Molnár S., Tanos P., Kovács J., Molnár M., Szabó I., Hatvani I.G. (2019): Estimation of dissolved oxygen in riverine ecosystems: Comparison of differently optimized neural networks. Ecological Engineering, Vol. 138, pp. 298-309.
2. **Csábrági A.,** Molnár S., Tanos P., Kovács J. (2017): Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. Ecological Engineering, Vol. 100, pp. 63-72. (ISSN: 0925-8574)
3. Molnár S., Molnár M., **Csábrági A.** (2014): Progress towards emission targets through the development of climate change policies and measures in Hungary. Időjárás - Quarterly Journal of the Hungarian Meterological Service, Vol. 118 (4), pp. 293-305.
4. **Csábrági A.,** Molnár S., Tanos P., Kovács J. (2015): Forecasting of dissolved oxygen in the river Danube using neural networks. Hungarian Agricultural Engineering, Vol. 27, pp. 38-41.
5. Molnár S., Somogyi F., **Csábrági A.** (2011): Comprehensive assessment of future energy needs and the role of alternative energy source. Hungarian Agricultural Engineering, 23, pp. 117-119.
6. Molnár S., Molnár M., **Csábrági A.** (2011): Impact assessment of mitigation strategies in the Hungarian agriculture. Journal of Agricultural Informatics, 2, pp. 10-17. (ISSN 2061-862X)

Proofread articles in Hungarian

1. Molnár S., **Csábrági A.** (2010a): Externális költségek vizsgálata az erőművi kibocsátások terén EcoSense modellel. Acta Agraria Kaposváriensis, 14(3), 69-77. o.
2. **Csábrági A.,** Molnár S., Tanos P., Kovács J. (2019): Neurális hálózatok alkalmazása ökológiai rendszerek vizsgálatában. Mezőgazdasági Technika, LX. évf. 3. sz., 1-5. o.
3. **Csábrági A.,** Molnár S., Tanos P., Kovács J. (2019): Neurális hálózatok alkalmazása hazai vízminőségi vizsgálatok során. Mezőgazdasági Technika LX. évf. 6. sz., 2-5. o.