



Estimation of organic matter dependent on different variables in drinking water network using artificial neural network and multiple regression methods

Sayıter YILDIZ^{1,*} , Can Bülent KARAKUŞ² 

¹Sivas Cumhuriyet University, Faculty of Engineering, Department of Environmental, Sivas/TURKEY

²Sivas Cumhuriyet University, Faculty of Architecture, Fine Arts and Design, Department of Urban And Regional Planning, Sivas/TURKEY

Abstract

The aim of this study is to estimate of organic matter values based on chlorine and turbidity values with the help of ANN and multiple regression (MR) methods. Three different models were done with ANN, and the statistical performance of these models was evaluated with statistical parameters like; μ , SE, σ , R^2 , RMSE and MAPE. The R^2 value of the selected best model was found to be quite high with 0.94. The relationship between the evaluation results of the ANN model and the empirical data ($R^2 = 0.92$) showed that the model was quite successful. In the MR analysis, R^2 was determined as 0.63, and a middling significant ($p < 0.05$) relationship was found. Since the calculated F value was greater than the tabulated F value, it was concluded that there is a clear relationship between dependent and independent variables. In addition, spatial distribution maps of chlorine, turbidity, organic matter values were created with the help of the GIS. With these maps, the estimated distribution of the measured parameters in the whole city network was accomplished. This study revealed that turbidity and chlorine parameters are related to organic matter value, and by establishing this relationship, organic matter can be estimated by ANN.

Article info

History:
Received: 15.03.2021
Accepted: 04.06.2021

Keywords:
Drinking water network,
Organic matter,
Artificial neural network,
Multiple regression methods.

1. Introduction

Water quality is a term describes the physical, chemical, and biological properties of water according to the suitability of water for a particular use. Water quality is influenced by natural and anthropogenic influences [1].

Natural organic matter is defined as a mixture of complex and diversified organic compounds resulted from natural processes occurring in the environment. Dead and live plants, animals, microorganisms, and their decomposition products can be precursors of natural organic matter [2]. Therefore, natural organic matter emerges as a result of contact between water present in the hydrological cycle and dead or living organic matter [3]. Natural organic matter in aquatic environments originates of both natural and human origin. However, the main source of natural organic matter is terrestrial vegetation and soils [4]. It is found widely in both surface and groundwater as a result of biological, geological, and hydrological interactions [5].

Organic matter in water is one of the most significant parameters affecting water quality. Organic substances found naturally in surface and underground water sources cause undesirable problems in many cases. The most important of these problems is that chlorine added to water for disinfection purposes creates trihalomethane (THM) compounds and other halogenated organic compounds as a result of the reaction with humic substances or other anthropogenic compounds in water [6, 7]. Since natural organic materials can cause problems with color and taste as well as forming disinfection by-products, the presence of natural organic matter in drinking water has attracted much interest in recent years [8].

Chlorine and chlorine compounds are the materials most commonly used in water treatment facilities. The most important feature of chlorine is that it has residual disinfection potential that prevents both the growth of microorganisms in drinking water networks and the entry of contaminants due to pipe breaks, maintenance of the network, negative pressure problems [9].

*Corresponding author. e-mail address: sayildiz@cumhuriyet.edu.tr.

The suspended particles or colloidal substances in the water cause turbidity that prevents the transmission of light in the water. Turbidity can be caused by organic or inorganic substances, or a combination of the two. Usually microorganisms (viruses, bacteria and protozoa) are added to the particles to remove turbidity by filtration, which greatly reduces the microbial pollution in the treated water [10].

Turbidity is a measure of the refractibility of light for water, and it has been used traditionally to indicate the quality of drinking water. Although microbiological contamination is usually joined by increased turbidity, other factors such as organic matter and silt also impact the turbidity levels in the water exiting the treatment plant [11]. Acceptable turbidity limits for water exiting from the treatment plant may vary between countries, but are generally less than 1 or 2 NTU [12]. However, to ensure the efficacy of disinfection, the turbidity must not be higher than NTU and Much less is preferred [10].

High levels of turbidity can stimulate bacterial growth by protecting microorganisms from the impacts of disinfection, which causes a great demand for chlorine. It is imperative to implement a comprehensive management strategy whereby multiple barriers are used in conjugation with disinfection to block or eliminate bacterial pollution, including water source protection and suitable treatment processes, in addition to protection during storage and distribution [10].

Water quality data are generally required to define the efficiency of water pollution control measures and the compliance level with determined standards of quality [13]. There is also a need for evaluating general water quality conditions and modeling water quality processes over a wide area. Therefore, water quality monitoring programs help to illuminate various processes that affect water quality and provide necessary information to water managers in decision-making [14].

In recent years, various studies have been carried out on water quality prediction models [15, 16]. However, traditional methods of data processing are no longer good enough to solve the trouble, as many factors that affect water quality have a complex nonlinear relationship [17]. On the other hand, ANN has been widely adopted for system identification, model definition, design optimization, and analysis, and

prediction, which can mimic the basic features of the human brain like self-adaptation, self-regulation, and fault tolerance [18, 19]. ANN networks can map the non-linear relationships that form the properties of aquatic ecosystems, and this distinguishes them from other statistical-based water quality models that suppose a linear relationship among response and prediction variables and their natural distribution [20]. In the last two decades, ANNs have made significant progress in practice in nearly all research areas [21-26].

In this study; Depending on the chlorine and turbidity amounts in the drinking water network of Sivas city, the amount of organic matter in the water was estimated by ANN and multiple regression (MR) methods. The predictability of this parameter, which is very important in terms of drinking water quality, depending on its varying values in different parts of the network was investigated. In addition, spatial distribution maps of chlorine, turbidity, organic matter measured in the drinking water network with the help of GIS and organic matter estimated by ANN were created. With these maps, the estimated distribution of the measured and predicted parameters in the whole city network was revealed. This study will contribute to different environmental studies where ANN, multiple regression analysis and spatial distribution maps will be evaluated together.

2. Materials and Methods

In this study, the results of turbidity, chlorine and organic matter analysis performed on samples taken from Sivas city drinking water distribution network by Sivas Municipality were used. The results of samples taken from 43 different points (Figure 1) that will represent the network starting from near the tank feeding the city network to the end point of the network were used in ANN and MR modeling. During the study, computer aided software program MATLAB R2013 was used for ANN calculations while Excel 2010 was utilized for regression analysis.

Using the IDW (Inverse Distance Weighted) interpolation method in the Spatial Analyst Module of the ArcGIS 10.2 software, spatial distribution maps were created for the measured turbidity, chlorine, organic matter values and the estimated organic matter values in the drinking water line in the study area.

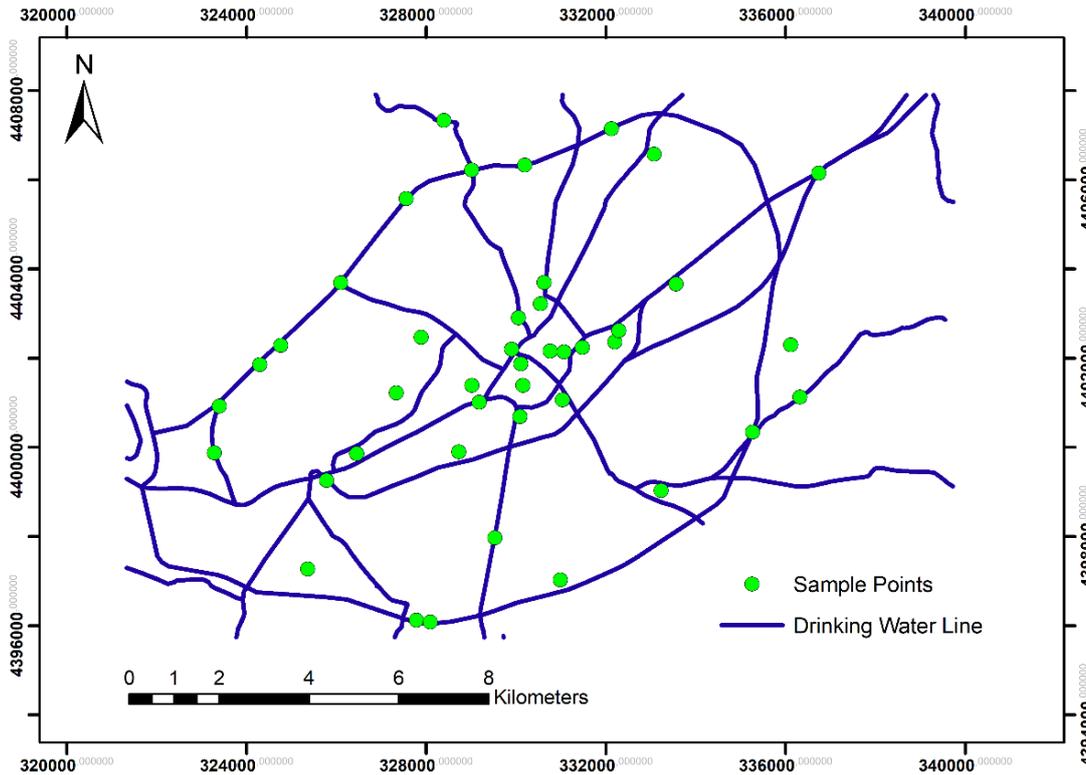


Figure 1. Sampled points on the network

2.1. Artificial Neural Network (ANN) Model

ANN was used to model the relationship between different data obtained by empirical methods and the estimation of a variable accordingly. In Figure 2, a simple ANN architecture was presented where the inputs are x_1, x_2, \dots, x_n and the weight coefficients of each input are $W_{k1}, W_{k2}, \dots, W_{kn}$. Here, x_n represents the input signals and W_{kn} represents the weight coefficients of these signals. The results from the thresholding function of the Y network are shown [25].

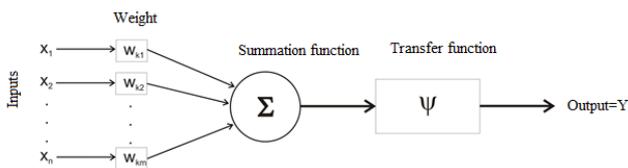


Figure 2. ANN cell model

The back propagation algorithm is a training algorithm and is widely used especially in engineering applications. The number of hidden layers in ANN can be augmented depending on nature of the problem [27]. The simple architecture of ANN's back propagation algorithm is given in Figure 3.

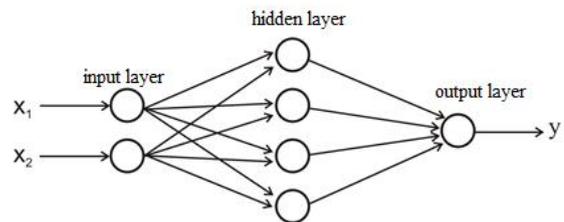


Figure 3. Simple architecture of ANN's back propagation algorithm

3. Results and Discussion

3.1. Organic matter estimation with ANN

The artificial neural network uses a modular neural network structure, which is a very strong computational technique, to model complex nonlinear relationships, especially when the relationship between variables is unknown in detail [28]. The basic structure of an ANN model generally consists of three various layers; The input layer, the hidden layer or layers and the output layer. The data is entered into the model and the weighted sum of the input is calculated in the input layer, The data is processed in the hidden layer or layers, while in the output layer the ANN results are produced. Each layer is made up of one or more fundamental components called a neuron or node [29].

The foundations of a neural network are the neurons that make up the hidden and output layers of the network (Fig.4).

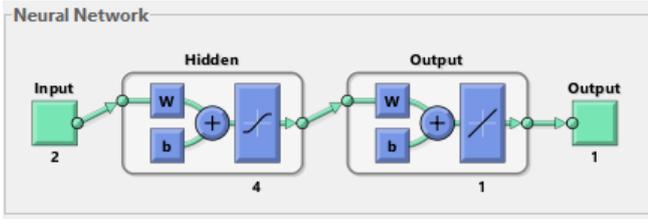


Figure 4. ANN structure

When designing an ANN, determining the number of neurons in the hidden layer is important. Too many hidden nodes can lead to over-compliance issues. Too few nodes in the hidden layer can cause non-compliance problems [30]. The number of hidden layers is chosen depending on the intricacy of the problem, yet one hidden layer is usually adequate to model most problems [31].

The training, validation and testing data of the ANN model that provides the best estimation are given in Figure 5. The statistical performance of the models was estimated depending on the statistical parameters μ , σ , R^2 .

In addition, RMSE and MAPE were used to evaluate the quality of models developed between the data estimated with ANN and the actual data [32]. RMSE is a measure of the quality of fit, and best describes the mean of the measurement error in predicting the dependent variable [33]. The statistical performance of the study is given in Table 1.

RMSE, MAPE and R2 are often used as a criterion to estimate network performance by comparing the error and measured data obtained from conjoint neural network studies [34]. RMSE and MAPE are calculated according to Equation 1-2.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (t_i - z_i)^2} \tag{1}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|(t_i - z_i)|}{z_i} \times 100 \tag{2}$$

Here; "t_i" and "z_i" are the estimated and actual outputs, while "n" represents the number of points in the data

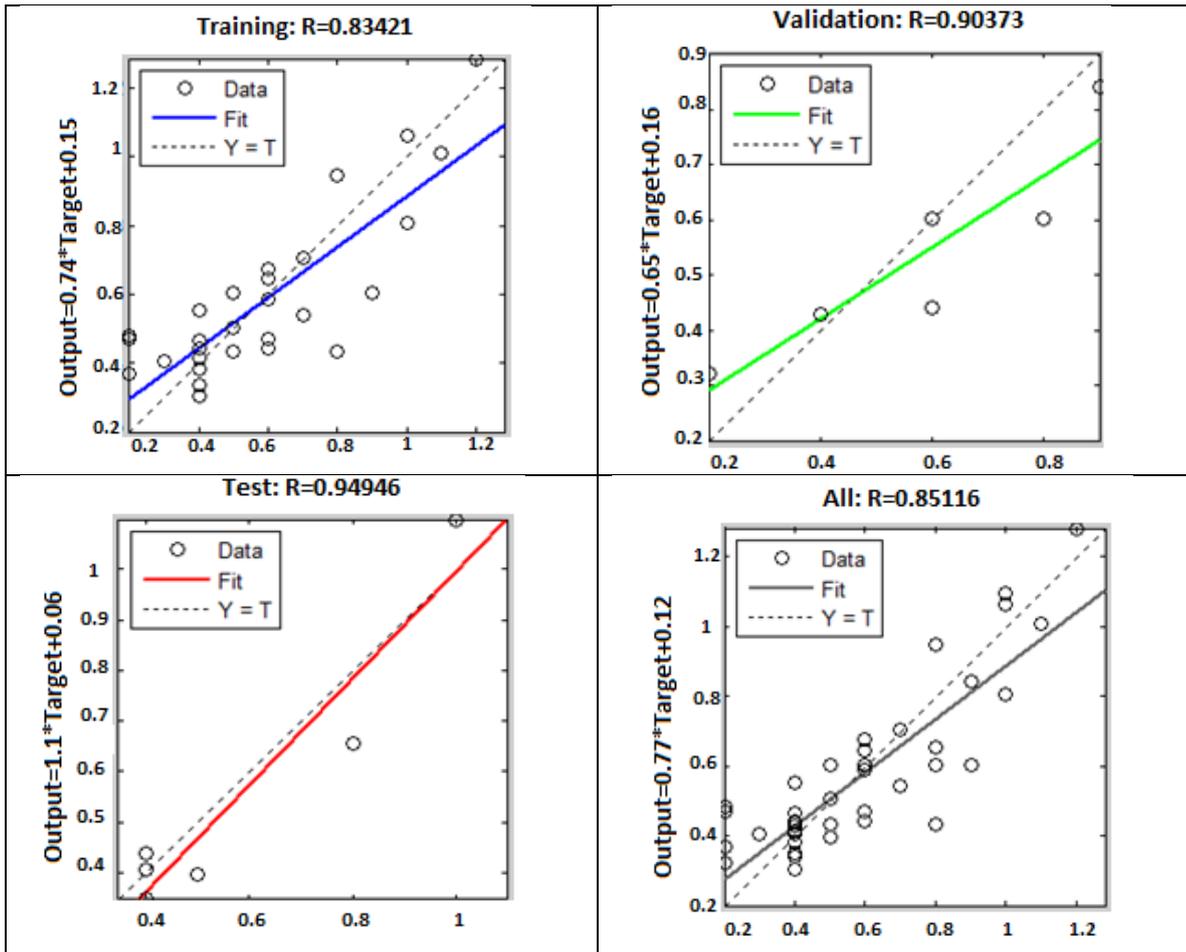


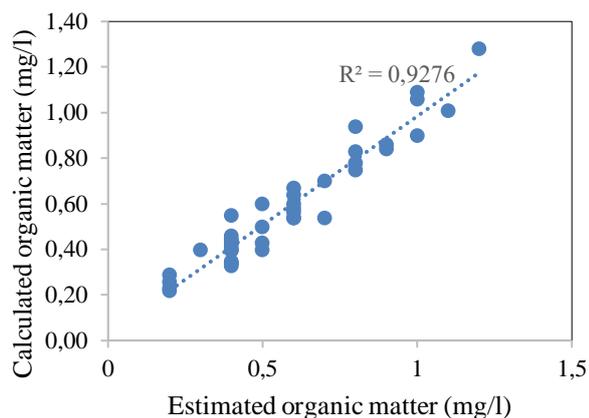
Figure 5. Training, validation and testing data obtained in ANN

Table 1. Statistical performance of ANN models

Model	Yapı	R ²	σ	SE	μ	RMSE	MAPE
I	2-2-1-1	0.82	0.22	0.10	1.01	0.61	19.42
II	2-3-1-1	0.90	0.14	0.07	0.96	0.62	13.10
III	2-4-1-1	0.92	0.13	0.06	0.99	0.62	11.61

As can be seen in Table 5, the results show that there is a considerable relationship between the values observed in the models created. However, the 2-4-1-1 model is seen as the best model in respect of R² and SE. RMSE values are very close to each other in all three models. The MAPE value is also an impartial statistic for measuring the predictive ability of a model. Low MAPE value indicates the best model performance [35].

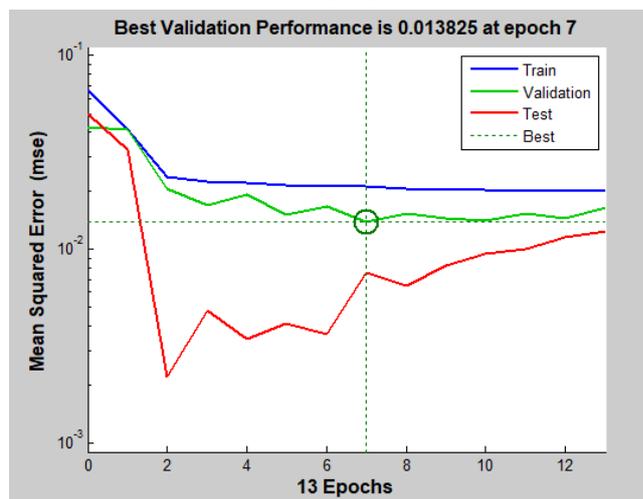
The relationship between the results of estimating the designed ANN model and the empirical data was arranged in order to assess the success of the ANN modeling used as an efficient tool. The diagram of the estimated organic matter values with ANN is shown in Figure 6.

**Figure 6.** Comparison of calculated and estimated organic matter results.

As seen in Figure 6, the estimated results and calculated data of the designed ANN model were compared, and it was found that they were in good harmony (R² 0.92).

ANN proved to be an efficient method for modeling organic materials with high R² values. The efficiency of the ANN model was settled based on maximizing R² and lowering the MSE value of the test set (1–13 neurons corresponding to the hidden layer). According to the graph of the lowest mean of squares error (MSE), and the number of epochs for the optimized ANN models (Fig.7), there was no significant change in the performance of the method after 7 stages. As seen in

Figure 6; the network was trained successfully with the algorithm of flexible back propagation.

**Figure 7.** Number of epochs for optimal ANN models according to MSE

3.2 Multiple regression analysis

The purpose of MR analysis is to define two or more independent variables at the same time to explain the variations of a dependent variable. In this study, turbidity and chlorine were accepted as independent variables, while organic matter was accepted as dependent variable. MR analysis was carried out to determine the relationship between organic matter and two independent variables. The studied statistical variables studied are given in Table 2.

Table 2. Statistical data of the variables

	Turbidity	Chlorine	Organic matter
Mean	0.57	0.29	0.57
Maximum	1.75	0.40	1.20
Minimum	0.30	0.15	0.20
Median	0.44	0.30	0.50
Variation	0.10	0.002	0.06
Standard Deviation	0.32	0.05	0.25
Skewness	2.23	-0.05	0.63
Kurtosis	4.68	0.19	-0.30

The approval of the model was made by considering the F test, t test and correlation coefficients. The statistical results of the model are given in Table 7. The importance of the R² value can be determined by means of the T-test, supposing that there is normal distribution of the variables and random selection of observations. The test compares the t value calculated using the null hypothesis to the tabulated t value. The

confidence level was chosen as 95% in this test, and the critical t value was obtained as ± 1.66. If the calculated t value is higher than the tabulated t value, the null hypothesis is inadmissible. This indicates R is important. If the calculated t value is lower than the tabulated t value, the null hypothesis is admissible. Thus, R is not important [24, 36]. Statistical results of the variables are given in Table 3.

Table 3. Statistical results

Independent Value	Dependent Value	R ²	Adjusted R ²	Unstandardized Coefficients	Standard Error	Calculated F value	Tabulated F value	Sign .
Turbidity	Organic matter	0.63	0.61	0.63	0.07	33.59	0.29	0.00
Chlorine				-0.20				
Independent Value	Calculated t		Tabulated t value		Sign .			
Turbidity	8.04		± 1.66		0.05			
Chlorine	-0.43				0.66			

As seen in Table 3, calculated t values are larger than tabulated t values. In this case, R is important. It was determined as R² = 0.63 and there is a middling significant (p<0.05) relationship. In addition, the calculated F value was higher than the tabulated F value. In this case, the null hypothesis is rejected. There is a real relationship between dependent and independent variables.

3.3. Spatial distribution maps

The Inverse Distance Weighted (IDW) Interpolation Method was used while creating spatial distribution maps of chlorine, turbidity, organic matter measured from the network and organic matter estimated by ANN.

Inverse Distance Weighted (IDW) Interpolation Method

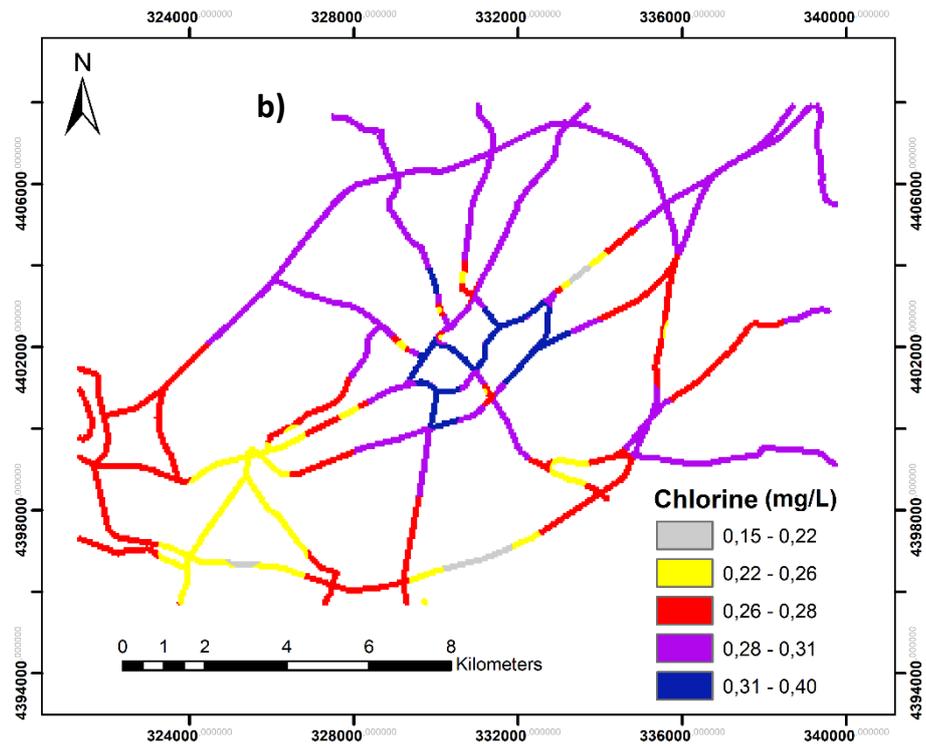
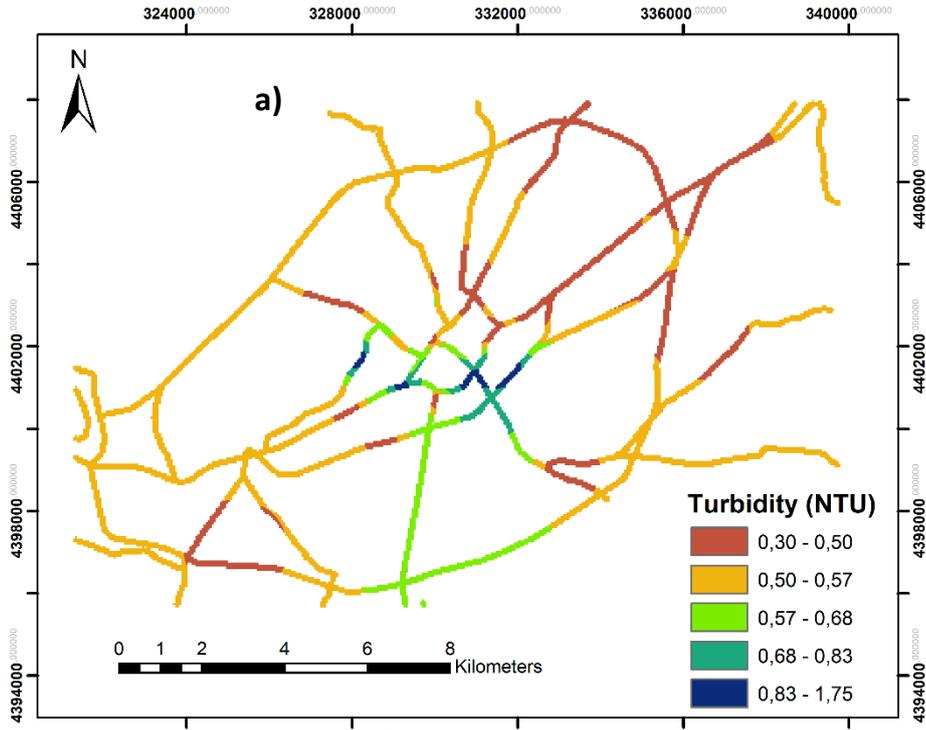
This method is an interpolation method that predicts values of cell by means of average values of sample data points in the vicinity of each cell. The sample points closest to the cell are given a high weight value. The further away from the estimation location, the effects of the points decrease. If any point is located in an area that is quite different from the estimation location, it may not be appropriate to consider a very distant point in this method. This problem can be solved provided that a sufficient number of points is taken into account and a surface is created for small areas. The number of points varies depending on the amount, distribution, and surface character of the

sample points [37]. The basis of this method is the calculation of distances from the desired point to data points, and the linear weighting of the effect of data points on the value at the desired point using an inverse function [38].

$$Z(X_o) = \frac{\sum_{i=1}^n Z(X_i) \cdot d_{i0}^{-r}}{\sum_{i=1}^n d_{i0}^{-r}} \tag{3}$$

Here; X_o is the position where the predictions are made, and this position is a function of adjacent measurements, n [Z (X_i) and i = 1, 2, ..., n]. r is the exponential number determining the assigned weight of each observation, and d is the distance between the observation position (X_i) and the estimated position (X_o).

The larger the exponent, the smaller the assigned weight of observations at a given distance from the estimation location. The increase in exponents shows that the estimations are very similar to the closest observations [39]. Spatial distribution maps created for each parameter are given in Figure 8.



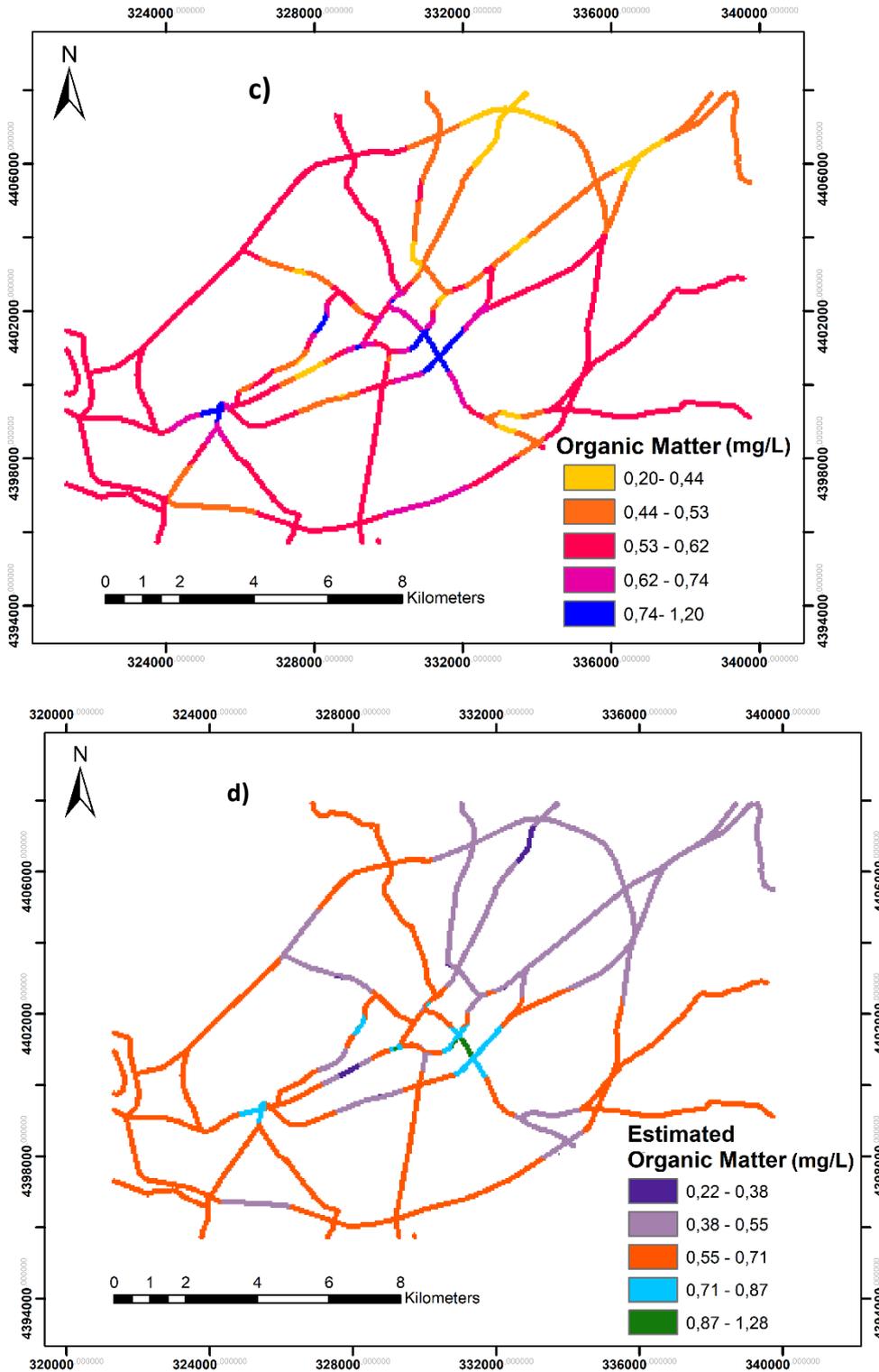


Figure 8. Spatial distribution maps for turbidity (a), chlorine (b), measured organic matter (c), estimated organic matter (d)

The measured turbidity values in the drinking water network vary between 0.30-1.75 NTU. The lowest values of turbidity in the drinking water network were observed in the north-east parts of the network, while the greatest values were seen in the middle parts of the network (Figure 8a). It was observed that chlorine and organic matter values were high in the middle parts of the city network (Figures 8b, 8c). When a general

evaluation is made in terms of measured and estimated organic matter in the city network; The measured and estimated organic matter amounts were approximately close to each other, and the max and min values of these amounts were observed in similar parts of the network (Figures 8c, 8d).

4. Conclusions

In this study, organic matter values were estimated by ANN and MR depending on the chlorine and turbidity measured in the drinking water network. Three different modelling were conducted with ANN, and the statistical performance of the models was evaluated with μ , SE, σ , R^2 , RMSE and MAPE parameters. The R^2 value of the graph of the empirical data and of the estimation results of the ANN model was 0.92 and showed that the model was quite successful. In the MR analysis, R^2 was determined as 0.63, and a middling significant ($p < 0.05$) relationship was found. Since the

calculated F value is higher than the tabulated F one, it is inferred that there is an actual relationship between dependent and independent variables. In addition, the estimated distributions of the measured parameters at 43 points representing the whole network gave an important prediction about the water quality in the network. In this study, it was revealed that the turbidity and chlorine parameters are related to the organic matter value and that the organic matter can be estimated by ANN by establishing this relationship.

Conflicts of interest

The author states no conflict of interests.

References

- [1] Khalil B., Ouarda T.B.M.J., St-Hilaire A., Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis, *Journal of Hydrology.*, 405 (2011) 277–287.
- [2] Chow C.W.K., Van Leeuwen J.A., Drikas M., Fabris R., Spark K.M., Page D.W., The impact of the character of natural organic matter in conventional treatment with alum, *Water Science and Tech.*, 40(9) (1999) 97–104.
- [3] Bridgeman J., Bierzoza M., Baker A., The application of fluorescence spectroscopy to organic matter characterisation in drinking water treatment, *Reviews in Environmental Science and Bio/Tech.*, 10(3) (2011) 277–290.
- [4] Matilainen A., Removal of the natural organic matter in the different stage of the drinking water treatment process, Thesis for the degree of PD, University of technology, (2007).
- [5] Sillanpää M., Ncibi M.C. Matilainen A., Vepsäläinen M., Removal of natural organic matter in drinking water treatment by coagulation: a comprehensive review, *Chemosphere*, 190 (2018) 54-71.
- [6] Fallahizadeh S., Neamati B., Fadaei A., Mengelizadeh N., Removal of Natural Organic Matter (NOM), Turbidity, and Color of surface water by integration of enhanced coagulation process and direct filtration, *Journal of Advances in Environmental Health Res.*, 5(2) (2017) 108-113.
- [7] Gümüş D., Akbal F., Removal of Natural Organic Matter In Drinking Waters And Prevention Of Trihalomethanes Formation, *Sigma: Journal of Engineering & Natural Sci.*, 31(4) (2013).
- [8] Trinh T.K., Kang L.S., Response surface methodological approach to optimize the coagulation-flocculation process in drinking water treatment, *Chem. Eng. Res. Des.*, 89(7) (2011) 1126-1135.
- [9] Ramos H. M., Loureiro D., Lopes A., Fernandes C., Covas D., Reis L.F., Cunha M.C., Evaluation of chlorine decay in drinking water systems for different flow conditions: from theory to practice, *Water Resources Managment*, 24(4) (2010) 815-834.
- [10] WHO, Guidelines for drinking water quality: training pack, 4nd ed. Geneva Switzerland, (2017) 564.
- [11] 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(1) (2007) 256.
- [12] Rouse R., New Drinking Water Regulations in the UK. London, Drinking Water Inspectorate, (2001).
- [13] Whitfield P., Goals and data collection designs for water quality monitoring, *Water Resour. Bull.*, 24 (4) (1988) 775–780.
- [14] Khalil B., Ouarda T.B.M.J., Statistical approaches used to assess and redesign surface water quality monitoring networks, *J. Environ. Monit.*, 11 (2009) 1915–1929.
- [15] Chen J.C., Chang N.B. Shieh W.K., Assessing wastewater reclamation potential by neural network model, *Eng. Appl. Artif. Intell.*, 16 (2003) 149–157.
- [16] Li R.Z., Advanced and trend analysis of theoretical methodology for water quality

- forecast, *J. Hefei Univ. Technol.*, 29 (2006) 26–30.
- [17] Xiang S.L., Liu Z.M., Ma L.P., Study of multivariate linear regression analysis model for ground water quality prediction, *Guizhou Sci.*, 24 (2006) 60–62.
- [18] Niu Z.G., Zhang H.W., Liu H.B., Application of neural network to prediction of coastal water quality, *J. Tianjin Polytechnic Univ.*, 25 (2006) 89–92.
- [19] Shu J., Using neural network model to predict water quality, *North Environ.*, 31 (2006) 44–46.
- [20] Lek S., Delacoste M., Baran P., Dimopoulos I., Lauga J., Aulagnier S., Application of neural networks to modelling nonlinear relationships in ecology, *Ecol. Model.*, 90 (1996) 39–52.
- [21] Chu W.C., Bose N.K., Speech signal prediction using feedforward neural network, *Electro. Lett.*, 34 (1998) 999–1001.
- [22] Messikh N., Samar M.H., Messikh L., Neural network analysis of liquid–liquid extraction of phenol from wastewater using TBP solvent, *Desalination.*, 208 (2007) 42–48.
- [23] Hanbay D., Turkoglu I., Demir Y., Prediction of wastewater treatment plant performance based on wavelet packet decomposition and neural networks, *Expert Syst. Appl.*, 34 (2008) 1038–1043.
- [24] Yıldız S., Değirmenci M., Estimation of oxygen exchange during treatment sludge composting through multiple regression and artificial neural networks, *International J. of Environmental Res.*, 9(4), (2015) 1173–1182.
- [25] Yıldız S., Artificial neural network (ANN) approach for modeling Zn (II) adsorption in batch process, *Korean J. of Chemical Eng.*, 34(9) (2017) 2423-2434.
- [26] Yıldız S., Artificial neural network (ANN) approach For modeling of Ni(II) adsorption from aqueous solution by peanut shell, *Ecol. Chem. Eng. S.*, 25(4) (2018) 581-604.
- [27] Demuth H., Beale M., Neural network toolbox for use with MATLAB, The MathWorks Inc. Natick, (2001) 840.
- [28] Smith M., Neural Networks for Statistical Modelling, Van Nostrand Reinhold, NY., (1994) 235.
- [29] Dreyfus G., Martinez J.M., Samuelides M., Gordon M.B., Badran F., Thiria S., Herault, Drinking Water and Health, Vol. 2., National Academy of Sciences, Washington, DC., (1980).
- [30] Shu C., Ouarda T.B.M.J., Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space, *Water Resour. Res.*, 43 (2007).
- [31] Rezvan K., Fakhri Y., Mehrorang G., Kheibar D., Back propagation artificial neural network and central composite design modeling of operational parameter impact for sunset yellow and azur (II) adsorption onto MWCNT and MWCNT-Pd-NPs: Isotherm and kinetic study, *Chemometrics and Intelligent Laboratory Systems.*, 159 (2016) 127–137.
- [32] Nabavi-Pelesaraei A., Kouchaki-Penchah H., Amid S., Modeling and optimization of CO₂ emissions for tangerine production using artificial neural networks and data envelopment analysis, *International Journal of Biosci.*, 4(7) (2014) 148–158.
- [33] Singh K.P., Basant A., Malik A., Jain G., Artificial neural network modeling of the river water quality-a case study, *Ecological Model.*, 220(6) (2009) 888-895.
- [34] Alves E.M., Rodrigues R.J., Dos Santos Corrêa C., Fidemann T., Rocha J.C., Buzzo J.L.L., et al., Use of ultraviolet–visible spectrophotometry associated with artificial neural networks as an alternative for determining the water quality index, *Env. Monit. and Assess.*, 190(6) (2018) 319.
- [35] Olyaie E., Banejad H., Chau K.W., Melesse A.M., A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: A case study in United States, *Env. Monit. and Assess.*, 187(4) (2015) 189.
- [36] Yıldız S., Karakuş C.B., Estimation of irrigation water quality index with development of an optimum model: a case study, *Environment, Development and Sustain.*, 22 (2020) 4771–4786.
- [37] Esri., Desktop spatial analysis for ArcGIS, Esri Information Systems Engineering and Education Ltd. Sti. 1st Edition, Ankara, (2014).
- [38] Loyd C.D., Local Models for Spatial Analysis, 2nd ed. ISBN 9780367864934, Temple University, Philadelphia, PA, USA., (2010) 98.

- [39] Aksu H.H., Hepdeniz K., Mapping with the aid of Geographic Information System and analysis of annual and monthly average maximum air temperature distribution in Burdur. *Mehmet Akif Ersoy University Journal of the Graduate School of Natural and Applied Sci.*, 7 (2016) 202-214.