Draw on artificial neural networks to assess and predict water quality

A Fernandes¹, H Chaves², R Lima³, J Neves³,⁴ and H Vicente¹,⁴,⁵

¹ Departamento de Química, Escola de Ciências e Tecnologia, REQUIMTE/LAQV, Universidade de Évora, Évora, Portugal
² Escola Superior Agrária de Beja, Instituto Politécnico de Beja, Beja, Portugal
³ Instituto Politécnico de Saúde do Norte, CESPU, Portugal
⁴ Centro Algoritmi, Universidade do Minho, Braga, Portugal
⁵ E-mail: hvicente@uevora.pt

Abstract. Water is one of the important vehicles for diseases of an infectious nature, which makes it essential to assess its quality. However, the assessment of water quality in reservoirs is a complex problem due to geographic limitations, sample collection and respective transport, the number of parameters to be studied and the financial resources spent to obtain analytical results. In addition, the period between sampling and analysis results must be added. This work describes the development of an Artificial Neural Network (ANN) to predict the biochemical and chemical oxygen demand based on the water pH value, the dissolved oxygen, the conductivity and its temperature. The models were trained and tested using experimental data (N=605) obtained from superficial water samples used to irrigate and produce water for public use, collected between September 2005 and December 2017. To evaluate the performance of the ANN models, the determination coefficient, the mean absolute error, the mean square error and the bias were calculated. It was determined that an ANN with topology 4-6-5-2 could be used successfully to predict the variables’ output. Indeed, good agreement was observed between the observed and predicted values, with the values of the coefficient of determination ranging from 0.813 to 0.979.

Keywords: Artificial Intelligence; Artificial Neural Networks; Biochemical Oxygen Demand; Chemical Oxygen Demand; Water Quality

1. Introduction

Water management is not only a quantitative problem that is embedded in its biogeochemical cycle, but also a qualitative one. Water is important to ensure the sustainability of life, and given its meaning for health, it is mandatory to ensure its quality. The quality assessment of water bodies implies the knowledge of a large set of parameters used to indicate its suitability for different purposes (e.g., drinking water production, irrigation). Water Quality (WQ) can be expressed by indicating the condition (i.e. dissolved or particulate) and the concentration of inorganic and organic substances present as well as some water-physical properties [1]. The WQ is assessed using analysis methods carried out in situ and in the laboratory. This process involves several steps, such as taking, preserving and transporting samples. In addition, it is still necessary to consider the time that has elapsed since the time of sampling and the completion of the laboratory analyzes as well as the financial resources expended. To maintain WQ in real time, developing computational models grounded on methodologies from Artificial Intelligence (AI) scientific area can be an alternative to WQ assessment. The Chemical Oxygen Demand...
(COD) and the Biochemical Oxygen Demand (BOD) enable to quantify the organic matter present in water. Some organic substances can be directly responsible for unpleasant colors and aromas or for the development of organisms from which they originate [2]. Organic matter can be of biological origin (e.g. proteins, fats and sugars) and/or due to anthropogenic activities. COD and BOD are criteria for controlling pollution and indicators of chemical and/or bacteriological contamination that can pose a threat to water consumers. COD can be defined as the amount of a particular oxidizer that reacts with the sample under controlled conditions. The BOD in turn measures the molecular oxygen consumption in a certain incubation period (five days) for the biochemical degradation of organic material [2]. This work aims the development of models grounded on AI techniques for problem solving, particularly ANNs [3], to predict BOD and COD based on pH, Dissolved Oxygen (DO), conductivity and water temperature.

In last decades several studies have been published demonstrating the effectiveness of ANNs in the prediction of surface water quality variables. In this context Maier and Dandy [4] present several case studies published in literature from 1992 to 1998. Chen et al. [5] compared the types and the use of ANNs taking into account the direction of the information flux, classifying them into feed-forward, recurrent and hybrid networks, and discussing the pros and cons of each different types of ANNs. Regarding the problems addressed, the prediction of salinity [4, 6], oxidability, total suspended solids [7], nitrate, manganese [8], dissolved oxygen [9–11], BOD [10] or COD [12] stand out.

This paper is divided into four sections. After the introduction, a new section is presented that describes the materials and methods used in this study. Section 3 shows how data is processed and the results obtained are discussed. The last section draws conclusions and outlines future work.

2. Materials and methods
This study was conducted at four dams located in the south of the district of Setúbal (southern Portugal). The Dams of Daroeira (37º 54’ 20.2'' N, 8º 19’ 27.0'' W), Fonte Serne (37º 52’ 54.0'' N, 8º 29’ 51.4'' W) and Campilhas (37º 50’ 37.4'' N, 8º 37’ 8.9'' W) are located at the municipality of Santiago do Cacém while Morgavel dam (37º 54’ 8.8'' N, 8º 45’ 49.3'' W) is in the municipality of Sines (Figure 1). Sampling was carried out between September 2005 and December 2017. Water Temperature, pH, DO, and Conductivity were evaluated in situ while BOD and COD were analyzed in lab.

![Figure 1. Geographical localization and distances between the dams where this study was conducted. A – Daroeira, B – Fonte Serne, C – Campilhas and D – Morgavel dams. Source: Google maps.](image)

2.1. Sample collection and preservation
The analytical control of water quality begins with sampling once a month (12 times a year) in the morning. The sampling procedures were conducted taking into account the recommendations presented in Standard Methods for the Examination of Water and Waste Water (SMEWW) [13]. For in situ determinations 50 ml wide-mouth polyethylene flasks were used, while 100 ml polyethylene flasks were used to collect the samples for COD analysis. Sulphuric acid (until pH<2) were added as preserving agent. Finally, for the BOD determination 1000 ml polyethylene flasks were used. The samples were...
transported in thermal bags containing ice accumulators. To ensure that the samples were not exposed to temperatures higher than those in the collection, a Data Logger, programmed to take minute-by-minute measurements, was placed in the pouch to record the temperature during transportation.

2.2. Analytical procedures
The water analyzes were carried out in a water laboratory accredited according to the ISO/IEC 17025. The pH, DO, Conductivity, Water Temperature, BOD and COD evaluation were carried out in accordance with SMEWW 4500-H† B, SMEWW 4500-O B, Portuguese version of European Standard 27888: 1996, SMEWW 2550 B, SMEWW 5210 B and SMEWW 5220 B [13].

2.3. Artificial neural networks
ANNs are computing tools designed to mimic the human brain and nervous system procedures. The MultiLayer Perceptron (MLP) is one of the most widespread ANNs architectures, in which neurons are disposed in layers, connected through forward connections [3]. In order to implement ANNs the software WEKA was used, maintaining unchanged the software set of parameters [14, 15]. Aiming to guarantee the statistical significance of the experiments, 25 replicates were realized. In each replicate the dataset was randomly split into two mutually exclusive sets. The training one, used to build the model, includes approximately two thirds of the cases. The remaining examples form the test set, used to assess the model performance.

3. Results and discussion

3.1. Database
The database contained 605 records and 6 fields (i.e., pH, Conductivity, Water Temperature, DO, COD, and BOD). Table 1 provides a statistical view of the variables recorded in the database. A review of Table 1 allows to state that all variables, except pH, exhibit large dispersion with variation coefficients that range from 25.4 to 84.8%. Such variability may be due to the large geographical area, differences in climate, seasonal effects and the numerous water sources. The pH, in turns, shows the lowest dispersion and this fact can be justified taking into account the buffering capacity of water bodies. Nonetheless, the values presented in Table 1 are in accordance with the ones reported in several similar studies [9, 10].

| Variable                                           | Minimum | Maximum | Mean  | Standard Deviation | Coefficient of Variation (%) |
|----------------------------------------------------|---------|---------|-------|--------------------|------------------------------|
| pH (Sørensen scale)                                | 6.5     | 9.9     | 7.8   | 0.7                | 9.0                          |
| Dissolved Oxygen (%)                               | 40.3    | 184.5   | 91.4  | 23.2               | 25.4                         |
| Conductivity (μS/cm)                               | 97.9    | 667.9   | 201.2 | 75.9               | 37.7                         |
| Water Temperature (°C)                             | 6.2     | 28.7    | 16.0  | 4.7                | 29.3                         |
| Biochemical Oxygen Demand (mg/dm³)                 | 0       | 13.4    | 2.9   | 2.4                | 84.8                         |
| Chemical Oxygen Demand (mg/dm³)                    | 9.2     | 48.9    | 19.4  | 6.1                | 31.3                         |

To evaluate the interrelationships between the variables, the Pearson correlation coefficients between the input variables (pH, DO, Water Temperature and Conductivity) and the output variables (COD and BOD) were computed (Table 2). A review of Table 2 reveals that the values of the coefficients are small and are all in the range from 0.08 … 0.43, which shows that there is no direct relationship among them.
Table 2. Pearson correlation coefficients among variables.

| Input Variables       | Output Variables | Biochemical Oxygen Demand (BOD) | Chemical Oxygen Demand (COD) |
|-----------------------|------------------|---------------------------------|------------------------------|
| pH                    |                  | 0.11                            | 0.21                         |
| Dissolved Oxygen      |                  | 0.14                            | 0.12                         |
| Conductivity          |                  | 0.08                            | 0.43                         |
| Water Temperature     |                  | 0.13                            | 0.32                         |

3.2. Artificial neural networks models

To achieve the best prediction of the variables COD and BOD, various network structures have been developed and evaluated. To assess the performance of the different networks, the Mean Absolute Error (MAE) and the Mean Square Error (MSE) were estimated. Such metrics compute mean errors but do not give any sign about the type of error. Thus, aiming to fulfil this lack of information, i.e., to estimate whether the COD and BOD values were overestimated or underestimated the average of all individual errors, known as \( \text{Bias} \), was computed. The results obtained are present in Table 3 for the networks tested. A review of Table 3 reveals that the 4-6-5-2 ANN minimizes \( \text{MAD} \), MSE and has a near zero \( \text{Bias} \) for both sets.

Table 3. Mean Absolute Error (MAE), Mean Square Error (MSE) and Bias for the ANN topologies tested.

| ANN Topology | Training Set | Test Set |
|--------------|--------------|----------|
|              | MAE\(^a\)    | MSE\(^b\) | Bias\(^c\) | MAE\(^a\) | MSE\(^b\) | Bias\(^c\) |
| 4-15-2       | 1.637        | 5.783     | 50.80      | 6.137      | 0.432     | -0.214     | 1.712      | 5.996      | 7.511     | 58.22      | -0.523     | -0.278     |
| 4-6-5-2      | 0.284        | 2.415     | 0.132      | 8.083      | -0.258    | 0.021     | 0.324      | 2.643      | 0.148     | 9.859      | 0.036      | 0.037      |
| 4-8-5-2      | 1.701        | 7.258     | 6.699      | 82.46      | 0.352     | -0.103    | 1.728      | 7.567      | 7.596     | 84.24      | -0.367     | 0.201      |
| 4-10-7-2     | 0.523        | 3.925     | 0.561      | 22.78      | -0.502    | 0.118     | 0.685      | 4.487      | 0.616     | 27.25      | 0.048      | -0.023     |
| 4-11-10-2    | 1.585        | 5.094     | 5.814      | 38.48      | 0.232     | 0.654     | 1.590      | 5.291      | 5.887     | 46.46      | 0.432      | -0.752     |
| 4-16-12-2    | 1.464        | 7.203     | 4.234      | 60.97      | -0.298    | -0.016    | 2.285      | 9.623      | 8.189     | 76.99      | -0.196     | 0.007      |

\(^a\) \( \text{MAE} = \frac{\sum_{i=1}^{N}|A_i - A'_i|}{N} \)

\(^b\) \( \text{MSE} = \frac{\sum_{i=1}^{N}(A_i - A'_i)^2}{N} \)

\(^c\) \( \text{Bias} = \frac{\sum_{i=1}^{N}|A_i - A'_i|}{N} \)

where \( A \) and \( A' \) stand for measured and predict values, while \( N \) indicates the number of cases.

The ANN, which was selected to predict COD and BOD based on pH, Water Temperature, Conductivity and Oxygen Content includes a four nodes input layer, six and five nodes hidden layers and an output layer with two nodes (Figure 2).

Figure 3 shows the diagrams between experimental and predicted BOD and COD values for training and test sets. The Determination Coefficient (\( R^2 \)) regarding training set were 0.975 and 0.813 for BOD and COD, respectively, while they were 0.979 and 0.822 for the test set. The accordance among measured and predict values for COD and BOD, \( R^2 \), MAE and MSE (Figure 3 and Table 2) allows to state that the model fits well to the data being useful to predict the COD and BOD.
Figure 2. Structure of the ANN selected to predict COD and BOD.

Figure 3. Graphical representation of ANN responses versus measured values regarding BOD and COD. (a) training data, (b) test data.

To reinforce the previous statements, Figure 4 presents the plot errors versus the ANN responses for both output variables and training and test sets. A review of Figure 4 allows to state that the relationship among errors and ANN responses shows complete independence and a random distribution for both COD and BOD. The previous statement is supported by the small values of the determination coefficients between 0.013 and 0.138. The points are well distributed above and below of the correct prediction line, i.e. the horizontal line of the ordinate zero. This type of display provides very useful information on model
adaptation. Thus, when the distribution of errors is random, it indicates that the model performs well. Conversely, if a non-random distribution is discernible, the model performance is not satisfactory [16].

3.3. Sensitivity Analysis
Sensitivity analysis is a process that can be used to assess the model response when the input values are changed. In the present study, the sensitivity based on variance [6, 17] was followed to evaluate the effects of each input variable on the ANN outputs. The outcomes are displayed in Figure 5 and suggest that ANN outputs are influenced in a similar way by the inputs, although the water temperature and pH have a slightly higher influence.

![Figure 4](image-url)
**Figure 4.** Graphical representation of errors versus ANN responses regarding BOD and COD. (a) training data, (b) test data.

![Figure 5](image-url)
**Figure 5.** Influence of each input variable on the ANN outputs.
4. Conclusions
This work presents an ANN model to forecast the BOD and the COD in superficial waters. Various ANNs architectures have been developed and evaluated using data collected from an accredited water laboratory in southern Portugal over a period of 17 years (from September 2005 to December 2017). The feedforward network with back propagation learning algorithm was used. The selected network exhibit a good performance in predicting of BOD and COD based on pH, Water Temperature, Conductivity and Oxygen levels and showed no overfitting. The promising results obtained in this study demonstrate the usefulness of ANNs as tools to predict water quality parameters, contributing to improve the management of the reservoirs as well as to preserve the quality of the water.

Acknowledgments
This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References
[1] Bagchi B 2013 Water in Biological and Chemical Processes (Cambridge: Cambridge University Press)
[2] Alley E R 2000 Water Quality Control Handbook (New York: McGraw-Hill)
[3] Haykin S 2009 Neural Networks and Learning Machines (New York: Prentice Hall)
[4] Maier H R and Dandy G C 2000 Application of artificial neural networks to forecasting of surface water quality variables: issues, applications and challenges Artif. Neur. Networks Hydrol. (Water Science and Technology Library) eds Govindaraju R S and Rao A R (Dordrecht: Springer) 136: 287–309
[5] Chen Y, Fang X, Yang L, Liu Y, Gong C and Di Y 2019 J. Phys.: Conf. Ser.1237 042051
[6] Maier H R and Dandy G C 1996 Water Resour. Res. 32 1013–22
[7] Vicente H, Couto C, Machado J, Abelha A and Neves J 2012 Int. J. Design Nature Ecodynamics 7 310–9
[8] Vicente H, Dias S, Fernandes A, Abelha A, Machado J and Neves J 2012 J. Water Supply Res. T. 61 446–59
[9] Palani S, Liong S-Y and Tkalich P 2008 Marine Poll. Bull. 56 1586–97
[10] Singh K, Basant A, Malik A and Jain G 2009 Ecol. Modelling 220 888–95
[11] Ahmed, A A M 2017 J. King Saud Uni. – Engin. Sci. 29 151–8
[12] Ruben G B, Zhang K, Bao H, Mao X 2018 Water Resour. Manage. 32 273–83
[13] Eaton A, Clesceri L, Rice E and Greenberg A 2017 Standard Methods for the Examination of Water and Wastewater (New York: American Public Health Association/Water Works Association/Water Environment Federation)
[14] Frank E, Hall M and Witten I H 2016 The WEKA Workbench – Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques” Morgan Kaufmann, 4th ed
[15] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten I H 2009 SIGKDD Explor. 11 10–8
[16] McBride G B 2005 Using Statistical Methods for Water Quality Management Issues, Problems and Solutions (Hoboken: John Wiley & Sons)
[17] Kewley R, Embrechts M and Breneman C 2000 IEEE Transac. Neural Networks 11 668–79