A study of water quality modelling with the artificial neural network method in Surabaya river

Riyanto Haribowo, Very Dermawan, Halita Fitrina
Water Resources Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, 65145, Indonesia
E-mail: riyanto_haribowo@ub.ac.id

Abstract. Decreasing river water quality due to the environmental conditions around the river can cause changes in the value of river benefits and endanger the environment. Thus, it becomes necessary to improve river water quality management by monitoring water quality. The purpose of this study is to model predictions of water quality conditions. The method is by analysing the parameters of water quality (BOD, COD, DO, pH, temperature) influenced by rainfall, catchment area, and land use. Data analysis was performed by the artificial neural network (ANN) method, using the Matlab R2014b software. In the process of network development, the most effective model was obtained with a variation of 75% training data with 5 hidden layers and a maximum epoch of 2000. From the results of the model, the relative errors (RE) for BOD, COD, DO, pH, and temperature were found to be 7.80%, 6.33%, 6.83%, 1.92% and 1.05% respectively, with an overall average of 4.79%. Because this RE value is quite small and less than 5%, it can be concluded that the artificial neural network method is quite effective for monitoring water quality conditions in the study area.

Keywords: water quality, Surabaya River, artificial neural network, Matlab R2014b.

1. Introduction
Increased population growth leads to an increased demand for land to live on, causing large amounts of green open land to be converted into settlements and industries. The existence of settlements and industries along the river can cause a decrease in water quality because the resulting waste is discharged directly into nearby rivers [10].

The decline in river water quality due to environmental conditions around a river can cause changes in the beneficial values of a river and endanger the environment. To reduce and anticipate this possibility, water quality management efforts need to be performed. One of the efforts to improve river water quality management is by water quality monitoring [3].

Water quality monitoring functions to provide basic information that can be used as a reference for determining water quality standards. This information can then be followed up to improve the existing water quality [3].

There are several methods that can be used to classify, predict, and simulate water quality. One of the methods that can be used is through Artificial Neural Networks or ANNs [4]. Artificial neural networks are designed to simulate the workings of the human brain. Artificial neural networks have the ability to recognize activities based on past or existing data. Existing data are studied in order to make decisions on data that have never been studied. This method is also able to be used to solve problems...
that are not structured and difficult to define. This reason underlies the use of the artificial neural network method to solve problems in various fields and disciplines [6].

In this study, ANN is applied to model the prediction of water quality conditions through the parameters of pH (acidity), DO (dissolved oxygen), BOD (biochemical oxygen demand), COD (chemical oxygen demand), and temperature in the downstream part of Surabaya River, specifically the monitoring point of Gunungsari Weir.

Previous studies had been conducted with the ANN method to predict water quality in Surabaya River. However, these were only based on rainfall and water quality data of the previous monitoring points and did not include the influence of the catchment area of each monitoring point and land use, as factors that affect water quality in Surabaya River [1, 2, 7, 11, 12]. Based on these studies, a modelling study needs to be conducted again in the Surabaya River by including the influence of rainfall, water quality of previous monitoring points, catchment area of each monitoring point, and land use.

2. Materials and Methods

2.1. Study Location

This study was conducted to predict water quality in Surabaya River, specifically at the monitoring point of Gunungsari Weir located in the City of Surabaya. The City of Surabaya has an area of approximately 1,191.25 km². Geographically, the City of Surabaya is located within the coordinates of 7°09’ SL to 7°21’ SL and 112°36’ EL to 112°54’ EL. The topography is flat land with elevations ranging from 0-20 meters above sea level.

![Figure 1. Map of the locations of observation points for water quality parameters. Source: Perum Jasa Tirta I (2018)](image)

In this study, water quality sampling was performed at 8 water quality monitoring points by Perum Jasa Tirta I (Figure 1). In addition to water quality, rainfall data were taken from 8 rain stations that are close to the water quality observation points.

2.2. Data

The following are the data needed for this study:
1. Map of the locations of water quality observation points.
2. Data of water quality (pH, COD, BOD, DO, and temperature) from 8 monitoring points over a period of 12 years.
3. Monthly rainfall data from 8 rain stations over a period of 12 years.
4. DEM map.

2.3. Analysis Method
This study involved the analysis of water quality parameters (BOD, COD, DO, pH, temperature) at the most downstream points of the Surabaya River, which are affected by rainfall, catchment area, and land use. Data analysis was performed by the Artificial Neural Network or ANN method using the Matlab R2014b software.

To determine the most optimal model for predicting water quality conditions at the Gunungsari Weir point, model calibration was performed by experimenting with various data variations [5]. The data variations among others included the use of 70% of data years or 8 data years (2006-2013), 75% of data years or 9 data years (2006-2014), and 80% data years or 10 data years (2006-2015) [8]. Each variation of data amounts was used for model training with 5 and 10 hidden layers and maximum epochs of 1000, 1500, and 2000. Relative error values were then calculated from the results of the training to obtain the smallest RE value. The most optimal model was determined based on the RE value.

Model verification was the process performed after the calibration phase was completed, which functions to test the performance of the model on data outside the calibration period [5]. After determining the most optimal model, verification was performed using the data years that were not used in the calibration.

This was then followed by making predictions using the generated data to determine the condition of water quality in the next 5 years. Results from the prediction stage were then compared with the values of Class II water quality as specified in Governor of East Java Regulation Number 61 of Year 2010 [9].

2.4. Relative Error (RE) Test
This test was utilized to determine the comparison between the value of one variable and the values of other variables, which is used as a benchmark for actual variables.

$$KR = \frac{\sum_{i=1}^{N}(P_i-Q_i)}{Y_1} \times 100\%$$  \hspace{1cm} (1)

Where:

- $P_i$ = observation data (actual data)
- $Q_i$ = forecast data (estimation result data)
- $N$ = the number of data points

3. Results and Discussion
This study required the areas of each catchment area from each water quality monitoring point as input data in the Artificial Neural Network method. The areas of each catchment area are needed to determine their effects on the condition of water quality, which shows the quantity of waste entering the river. However, this does not guarantee that a larger catchment area means a greater amount of waste entering the river.

The areas of each catchment area in this study were determined by processing of the DEM (Digital Elevation Model) map with the assist of the Arcmap 10.4 software. The total area of all catchment areas in the study location is 105,959.004 Ha. The areas of the catchment area of each monitoring point can be seen in Table 1.

After determining the areas of each catchment, digitization was then performed based on satellite imagery data available on Arcmap 10.4 to determine land use. Land use was categorized as four regions, which are industry, settlement, agriculture, and other (which include vacant land, shrubs, and basins). From the digitization results, the areas of each region were obtained. Next, the areas were accumulated for each point.

Based on the digitization results, it could be seen that the catchment areas for the points of Canggu Tambangan and Perning Bridge are dominated by agricultural land. Meanwhile, Jrebeng Bridge is
dominated by land for settlements, with almost the same area for agricultural land. Cangkir Tambangan and Bambe Tambangan are also dominated by agricultural land. For the catchment areas of Karangpilang and Sepanjang Bridge monitoring points, land use is dominated by settlements. The monitoring point of Gunungsari Weir is dominated by agricultural land. From Figure 2 it can be seen that agricultural land dominates overall land use and the rest is dominated by residential areas.

Table 1. Percentage of Land Use Areas for Each Monitoring Point

| No | Outlet Point      | Industry Area (%) | Settlement Area (%) | Agriculture Area (%) | Other Area (%) |
|----|-------------------|-------------------|---------------------|----------------------|---------------|
| 1  | Canggu Tambangan | 0.40              | 15.07               | 84.49                | 0.04          |
| 2  | Perning Bridge   | 0.92              | 13.28               | 85.67                | 0.12          |
| 3  | Jrebang Bridge   | 0.40              | 57.11               | 42.49                | 0             |
| 4  | Cangkir Tambangan| 5.08              | 19.40               | 75.08                | 0.44          |
| 5  | Bambe Tambangan  | 12.53             | 23.25               | 64.09                | 0.13          |
| 6  | Karangpilang     | 40.58             | 23.91               | 34.74                | 0.77          |
| 7  | Sepanjang Bridge | 14.12             | 58.60               | 26.70                | 0.59          |
| 8  | Gunungsari Weir  | 3.81              | 62.65               | 32.42                | 1.12          |

[Figure 2. Map of Surabaya River land use]

In determining the model using Matlab R2014b, the first step involved the development of network architecture to be used by determining several elements, which included network type, training function, function adaptation, performance function, and transfer function. In this study, the utilized type of network was feed-forward back propagation. The utilized training function was the function of TRAINCGF (Conjugate Gradient Backpropagation with Fletcher-Reeves Restarts). Adaptation of the learning function used LEARNGD, while the performance function used MSE (Mean-Square Error). The input transfer function used TANSIG and PURELIN for the last layer.
3.1. Model Calibration

3.1.1. 70% Data Variation
Based on Table 2, it is known that the value of R is quite good because it approached the number 1, where a greater R value leads to a smaller average RE value, indicating that the results of the training output are becoming better. From Table 2, it is known that the best results were obtained with a maximum epoch of 1500, with an R value of 0.997 and a mean or average RE of 5.56%.

Based on Table 3, it is known that the value of R is quite good because it approached the number 1, where a greater R value leads to a smaller average RE value, indicating that the results of the training output are becoming better. From Table 3, it is known that the best results were obtained with a maximum epoch of 1000, with an R value of 0.997 and a mean RE of 8.11%.

| Variation | Epoch | R    | Average of RE (%) |
|-----------|-------|------|-------------------|
| 70%       | 1000  | 0.98955 | 5.88               |
|           | 1500  | 0.99739 | 5.56               |
|           | 2000  | 0.98955 | 5.88               |

3.1.2. 75% Data Variation
Based on Table 4, it is known that the value of R is quite good because it approached the number 1, where a greater R value leads to a smaller mean RE value, showing that the results of the training output are becoming better. From Table 4, it is known that the best results were obtained with a maximum epoch of 2000, with an R value of 0.99531 and a mean RE of 4.79%.

| Variation | Epoch | R    | Average of RE (%) |
|-----------|-------|------|-------------------|
| 75%       | 1000  | 0.9995 | 7.40               |
|           | 1500  | 0.995  | 4.85               |
|           | 2000  | 0.995  | 4.79               |

Based on Table 5, it is known that the value of R is quite good because it approached the number 1, where a greater R value leads to a smaller mean RE value, showing that the results of the training output are becoming better. From Table 5, it is known that the best results were obtained with a maximum epoch of 1000, with an R value of 0.996 and a mean RE of 8.17%.

| Variation | Epoch | R    | Average of RE (%) |
|-----------|-------|------|-------------------|
| 75%       | 1000  | 0.996 | 8.17               |
|           | 1500  | 0.980 | 14.68              |
|           | 2000  | 0.984 | 13.90              |

3.1.3. 80% Data Variation
Based on Table 6, it is known that the value of R is quite good because it approached number 1, where a greater R value that leads to a smaller mean RE value shows that the results of training output are

| Variation | Epoch | R    | Average of RE (%) |
|-----------|-------|------|-------------------|
| 80%       | 1000  | 0.996 | 8.17               |
|           | 1500  | 0.984 | 13.90              |

Based on Table 6, it is known that the value of R is quite good because it approached number 1, where a greater R value that leads to a smaller mean RE value shows that the results of training output are becoming better.
becoming better. From Table 6, it is known that the best results were obtained with a maximum epoch of 1000, with an R value of 0.994 and a mean RE of 8.39%.

Based on Table 7, it is known that the value of R is quite good because it approached number 1, where a greater R value that leads to a smaller mean RE value shows that the results of training output are becoming better. From Table 7, it is known that the best results were obtained with a maximum epoch of 1500, with an R value of 0.988 and a mean RE of 12.85%.

Table 6. The Value of R and Average of Relative Error (RE) with the 80% Data Variation; 5 Hidden Layers for Epoch 1000, 1500, and 2000.

| Variation | Epoch | R    | Average of RE (%) |
|-----------|-------|------|-------------------|
| 80%       | 1000  | 0.994| 8.39              |
|           | 5     | 1500 | 0.988             |
|           |       | 2000 | 0.989             |

Table 7. The Value of R and Average of Relative Error (RE) with the 80% Data Variation; 10 Hidden Layers for Epoch 1000, 1500, and 2000.

| Variation | Epoch | R    | Average of RE (%) |
|-----------|-------|------|-------------------|
| 80%       | 1000  | 0.988| 12.90             |
|           | 5     | 1500 | 0.988             |
|           |       | 2000 | 0.982             |

3.2. Relative Error (RE)
To find out which model is the most optimal, relative error (RE) was calculated for each model that had been trained. After finding the RE for each parameter, the average or mean RE values were calculated. These averages were then compared to find the smallest average or mean RE value.

Table 8. Summary of Relative Error (RE) Calculation from the Results of the ANN Model

| Variation | Epoch | BOD  | COD  | DO   | pH   | Temp.  | Mean of RE (%) |
|-----------|-------|------|------|------|------|--------|----------------|
| 70%       | 1000  | 10.24| 8.12 | 8.12 | 2.04 | 0.90   | 5.88           |
|           | 5     | 9.82 | 5.83 | 9.08 | 2.05 | 0.99   | 5.56           |
|           | 2000  | 10.24| 8.12 | 8.12 | 2.04 | 0.90   | 5.88           |
|           | 10    | 14.99| 7.76 | 13.69| 2.66 | 1.48   | 8.11           |
|           | 1500  | 18.62| 8.11 | 16.32| 3.82 | 1.85   | 9.75           |
|           | 2000  | 15.57| 13.30| 15.33| 3.12 | 1.76   | 9.86           |
| 75%       | 1000  | 12.97| 10.25| 10.45| 1.95 | 1.36   | 7.40           |
|           | 5     | 8.04 | 6.39 | 6.85 | 1.93 | 1.06   | 4.85           |
|           | 2000  | 7.81 | 6.33 | 6.83 | 1.92 | 1.05   | 4.79           |
|           | 10    | 13.64| 8.78 | 13.96| 3.04 | 1.43   | 8.17           |
|           | 1500  | 28.04| 16.43| 21.85| 4.37 | 2.73   | 14.68          |
|           | 2000  | 27.33| 13.26| 21.33| 4.94 | 2.66   | 13.90          |
| 80%       | 1000  | 15.80| 10.34| 11.54| 2.82 | 1.48   | 8.39           |
|           | 5     | 26.72| 13.06| 23.17| 4.77 | 2.35   | 14.02          |
|           | 2000  | 26.64| 12.77| 22.56| 4.72 | 2.33   | 13.81          |
|           | 10    | 25.96| 16.27| 15.82| 3.98 | 2.47   | 12.90          |
|           | 1500  | 26.30| 15.90| 15.61| 3.98 | 2.47   | 12.85          |
|           | 2000  | 35.48| 20.28| 23.47| 4.31 | 2.57   | 17.22          |

Based on Table 8, the model with 75% data variation, 5 hidden layers, and a maximum epoch of 2000 had the most optimal results because the average generated RE value is less than 5%, specifically
being 4.97%. The BOD, OD, DO, pH, and temperature parameter mean values for the model are 7.81%, 6.33%, 6.83%, 1.92%, and 1.05%, respectively. This became the selected model for this study.

3.3. Model Verification
Model verification was performed using data that was not used for the calibration stage. Because the calibration data that were used for the selected model were the data from 2006-2014 (9 years), in the verification stage, the data that were used were the data from 2015-2017 (3 years). Verification was performed by testing the data using the parameters from the selected model, such as the number of hidden layers and the maximum number of epochs with the same treatment. In this verification stage, the target was reached at iteration 312 after 14 seconds, with validation check at 312 and step size having been fulfilled. The MSE value at epoch 312 was $5.76 \times 10^{-12}$, and the gradient at epoch 312 was $3.42 \times 10^{-8}$. The best performance on this network occurred at epoch 0, where MSE values reached $1.87 \times 10^{-9}$. Even though it did not achieve the desired performance, the network was still successful because the MSE value was still relatively small. For this network, epoch 312 showed a fairly good gradient value that was still above the minimum gradient value, being $3.42 \times 10^{-9}$.

![Figure 3. Regression plotting for data verification](image)

The results of the regression plotting showed quite good results, with the regression value for the overall data closely approaching the number 1, being 0.999 (Figure 3). After the training process was performed, the relative error was calculated. Relative error was calculated to find out the reliability value.

The average RE value of each water quality parameter were quite good, being less than 10%, with the BOD, COD, DO, pH, and temperature parameter values being 10.39%, 2.26%, 7.53%, 2.60%, and 0.92% respectively. Meanwhile, the average RE value is relatively small and close to the RE value in the calibration of the selected model, being 4.85%. This shows that the selected model is quite good to be used for predictions.

3.4. Water Quality Prediction
After obtaining an optimal model for predicting water quality in the downstream part of the Surabaya River, a prediction was made on the water quality in the next 5 years. This prediction was made to find out the condition of the water in Surabaya River for the period from 2018-2022. This prediction was
made by entering all the data from 2006-2022. Next, model training was performed using the parameters from the selected model with the same treatment.

After obtaining the prediction results for water quality at the monitoring point of Gunungsari Weir for 2018-2022 using the ANN method, the values that were obtained were then compared to the provisional values for Class II water quality criteria based on Government Regulation of the Republic of Indonesia Number 82 of 2001. This was performed to find out whether the water quality of the Surabaya River is still in accordance with the provisions of Governor of East Java Regulation No. 61 of 2010.

Figure 4. Graph of the BOD parameter for the Gunungsari Weir point for the 2018-2022 period

Figure 5. Graph of the COD parameter for the Gunungsari Weir point for the 2018-2022 period

Figure 6. Graph of the DO parameter for the Gunungsari Weir point for the 2018-2022 period
Figure 7. Graph of the pH parameter for the Gunungsari Weir point for the 2018-2022 period

Figure 8. Graph of the Temperature parameter for the Gunungsari Weir point for the 2018-2022 period

Based on the graph for the BOD parameter, many values are greater than the specified maximum limit of 3 mg/L. The average of the BOD parameter is also greater than the maximum limit of 4.42 mg/L, which is close to the maximum value for class III, being 6 mg/L (Figure 4). For the COD parameter, the resulting predicted value is close to the specified maximum limit of 25 mg/L, and the average value of the COD parameter is 22.18 mg/L, still below the maximum limit value (Figure 5). For the DO parameter, the average value is close to the minimum limit, being 3.76 mg/L (Figure 6). For the pH parameter, some values are outside the specified value range, but the average value is still within the specified range, being 6.37 (Figure 7). As for the temperature parameter, the resulting predicted values are still mostly at deviation 3, being not too deviant and the average being 29.54°C (Figure 8).

Looking at the obtained results, some parameters were not in accordance with the criteria for Class II water quality but were still close to the specified value. Similarly, there are several parameters that were still in accordance with the specified values.

4. Conclusion

Based on the results of the analysis and discussion, it can be concluded that the most optimal model is the model with the calibration parameters of 75% data used out of a total 12 years, 5 hidden layers, a maximum epoch of 2000, the TRAINCGF training function, and an overall RE value of 4.79%. After conducting verification of the selected model using data from the years that were not used, the obtained overall RE value is 4.85%. The prediction of water quality in the next 5 years showed that the water quality conditions, according to the parameters of COD, pH, and temperature, are still within the criteria of Class II water quality. For the BOD and DO parameters, the values are inferior to the specified criteria, but the values do not deviate too far.
References

[1] Badan Lingkungan Hidup. 2011. Status Lingkungan Hidup Daerah Provinsi Jawa Timur Tahun 2010 [Regional Environmental Status of the Province of East Java] (Surabaya: BLH Jawa Timur).

[2] Delio R. 2017. Aplikasi Jaringan Syaraf Tiruan untuk Memprediksi Kualitas Air Sungai di Titik Karangpilang Kota Surabaya [The Application of Artificial Neural Networks to Predict River Water Quality at Karangpilang, City of Surabaya] (Malang: Universitas Brawijaya).

[3] Gubernur Jawa Timur. 2010. Peraturan Gubernur Jawa Timur Nomor 61 Tahun 2010 tentang Penetapan Kualitas Air pada Air Sungai [Governor of East Java Regulation Number 61 of 2010 on the Establishment of Water Quality for River Water] (Surabaya: Pemerintahan Provinsi Jawa Timur).

[4] Haribowo R, Dermawan V, Yudha N. 2018. Application of Artificial Neural Network for Defining the Water Quality in the River Civil and Environmental Science Journal 1(1) 12-18.

[5] Harto S. 2009. Hidrologi Teori, Masalah, Penyelesaian [Hydrology: Theories, Issues, Resolutions] (Yogyakarta: Nafiri Offset) p 256.

[6] Hermawan A. 2006. Jaringan Syaraf Tiruan Teori dan Aplikasi [Artificial Neural Networks: Theories and Applications] (Yogyakarta: C.V Andi Offset).

[7] Ivantoro B A. 2017. Aplikasi Jaringan Syaraf Tiruan untuk Memprediksi Kualitas Air Sungai di Titik Bendung Gunungsari Kota Surabaya [The Application of Artificial Neural Networks to Predict River Water Quality at Gunungsari Dam, City of Surabaya] (Malang: Universitas Brawijaya).

[8] Maru’ao D O. 2010. Neural Network Implementation in Foreign Currency Exchange Prediction UG Journal 2(5) 19-38.

[9] Pemerintah Gubernur Jawa Timur. 2010. Peraturan Gubernur Jawa Timur Nomor 61 Tahun 2010 tentang Penetapan Kelas Air Pada Air Sungai. [Governor of East Java Regulation Number 61 of 2010 on the Establishment of Water Quality for River Water]

[10] Suwari, Riani E, Pramudya B, Djuwita I. 2010. Penentuan Status Mutu Air Kali Surabaya dengan Metode Storet dan Indeks Pencemaran [Determining the Water Quality Status of Surabaya River with the Storet and Pollution Index Methods] Majalah Ilmiah WIDYA 297 59-64.

[11] Winanto P D. 2017. Aplikasi Jaringan Syaraf Tiruan untuk Memprediksi Kualitas Air Sungai di Titik Bambe Tambangan Kabupaten Gresik [The Application of Artificial Neural Networks to Predict River Water Quality at Bambe Tambangan, Gresik Regency] (Malang: Universitas Brawijaya).

[12] Yudha N S. 2017. Aplikasi Jaringan Syaraf Tiruan untuk Memprediksi Kualitas Air Sungai di Titik Jembatan Jrebeng Kabupaten Gresik [The Application of Artificial Neural Networks to Predict River Water Quality at Jrebeng Bridge, Gresik Regency] (Malang: Universitas Brawijaya).