Hydroecological data recovery using artificial intelligence

Elvira Nafikova¹, Dmitriy Aleksandrov¹, Alsu Shaniyazova¹*, and Christina Bondar¹

¹Ufa University of Science and Technology, 32, Zaki Walidi street, Ufa, 450076. Russia

Abstract. Artificial neural networks and genetic algorithms have been tested for the restoration of missed hydroecological indicators (hydrological water quality parameters and bottom sediment quality parameters). Algorithms have been developed for recovering missing hydroecological data in the presence and absence of observation data from points upstream and downstream. Neural network models for the restoration of water quality indicators and bottom sediments were tested on the example of the catchment basin of the Bayda and Kidysh rivers in the territory of the Republic of Bashkortostan.

1 Introduction

When studying the hydrological processes of a watercourse, certain difficulties arise associated with data gaps and the selection of significant input parameters. Long-term series of accumulated observation data on water quality and the quality of bottom sediments of small rivers are characterized by the presence of gaps associated, for example, with the impossibility of observation, the complexity of the analysis of samples, loss of data, etc. As a rule, the density of gaps is high, and their location is irregular.

To restore the missing hydrometeorological data, statistical methods are widely used [1, 2], while the gaps are replaced by either the average value of this indicator or the missing data is removed from the time series.

The use of statistical methods leads to the loss of information or its significant distortion, which can subsequently lead to an incorrect forecast. In addition, the changing anthropogenic load on the water body makes it impossible to use these methods to assess changing hydrological and meteorological parameters. Therefore, statistical methods for forecasting hydrometeorological parameters are more widely used to calibrate other more complex forecasting methods [3, 4].

Researchers [5-8] propose hybrid models for predicting water flow in a watercourse based on the integration of various forecasting models and the selection of a large number of significant parameters that affect river flow and water quality.

In [9, 10], methods of hydrological analogy with the already studied territory or historical period are proposed for the restoration of hydrological data. The complexity of this method lies in the selection of an analogue river that would be subject to the same anthropogenic load as the studied watercourse.

* Corresponding author: ashaniyazova@bk.ru
In some works [11-13], attempts have been made to combine genetic algorithms (GA) and ANN, the possibilities of using GA to optimize the learning process of ANN, aimed at reducing the amount of computations, have been investigated, provided that the accuracy of the solution is maintained at the required level.

The analysis of works [11-13] showed that for the restoration of missing data of natural processes it is most preferable to use the ANN of the following structures: multilayer perceptron (MLP), radially basic networks (RBF), generalized regression networks (GRNN).

2. Materials and methods

In this work, a time series forecasting model using artificial neural networks (ANN) was used to recover the missing hydrological data.

A diagram of the structure of the neural network for the restoration of the indicators of water quality and quality of bottom sediments is shown in Figure 1.

Fig. 1. Structure of the neural network for the restoration of hydroecological data.

The authors, based on the ANN apparatus, carried out the restoration of the missing data in four stages:

- Preparation of initial data to restore the required indicator.
- Using trained networks for data recovery.

At the same time, it was proposed to use two methods for recovering missing data using artificial neural networks:

- Missing data recovery only in the presence of data of the recoverable indicator for a multi-year period (e.g., content of Cu in water).
- Missing data recovery in the presence of data of the recoverable indicator and analogous observation points for a multi-year period (e.g., the content of heavy metals in the water during the study period with the presence of gaps and indicators of water quality).

To restore the missing hydrological parameters, a model was used that analyzes the relationship between daily data on a hydrological indicator, which is a time series model (Figure 2). The use of such a model is based on the fact that real hydrological parameters are the result of the impact of all factors, including those that cannot be taken into account.
or described in numerical form. This model can only be implemented with significant amounts of data (i.e., the data must be daily and for a significant period).

![Input and output parameters for forecasting hydrological parameters.](image)

<table>
<thead>
<tr>
<th>Date</th>
<th>Sulfate level</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.05.2000</td>
<td>284</td>
</tr>
<tr>
<td>19.07.2000</td>
<td>183</td>
</tr>
<tr>
<td>05.09.2000</td>
<td>310</td>
</tr>
<tr>
<td>19.08.2003</td>
<td>457</td>
</tr>
<tr>
<td>02.12.2003</td>
<td>2220</td>
</tr>
<tr>
<td>31.05.2005</td>
<td>185</td>
</tr>
<tr>
<td>25.10.2005</td>
<td></td>
</tr>
<tr>
<td>09.06.2006</td>
<td>103</td>
</tr>
<tr>
<td>12.09.2006</td>
<td>1284</td>
</tr>
<tr>
<td>02.12.2007</td>
<td></td>
</tr>
<tr>
<td>04.06.2015</td>
<td>449</td>
</tr>
<tr>
<td>21.08.15</td>
<td>340</td>
</tr>
<tr>
<td>13.12.15</td>
<td></td>
</tr>
<tr>
<td>10.02.16</td>
<td>609</td>
</tr>
<tr>
<td>01.04.16</td>
<td>237.6</td>
</tr>
<tr>
<td>31.05.16</td>
<td>192</td>
</tr>
</tbody>
</table>

The implementation of the restoration of missed geoeological parameters using artificial neural networks was carried out in the Statistica 12.0 software product.

The scheme for recovering missing data in the presence of data of only the recoverable indicator for a long-term period is graphically interpreted in Figure 3.

When restoring the missing data with the presence of data of only the recoverable indicator, the input parameters of the ANN are the data of the observation point for a long-term period, the readings of which must be restored (hereinafter - the restored observation point).
Fig. 3. Scheme for recovering missing data in the presence of data of only the recoverable indicator for a long-term period.
When preparing the input data for the recovery model using ANN, the quality of the prepared initial data plays an important role:

- Sufficiency of the period of the initial data used: the period of the used data has to be at least 3 years to restore daily indicators with a gap density of 20% (accordingly, with an increase in the gap density, the required period of the initial data increases); at least 30 years for average long-term (taking into account 15-year and 11-year cycles of hydrological and meteorological processes);
- Use of data for a period with a stable anthropogenic load.

The ANN was trained based on the input data, adjusting the network weights and choosing the best ANN model to restore the required indicator.

The scheme for recovering missing data using artificial neural networks in the presence of data of the recovered indicator and data of analogous points for a long-term period is graphically interpreted in Figure 4.

Observation points located in similar physical and geographical conditions and/or the nearest observation points were considered as points-analogs in this work, for example, for river sections they were sections upstream and downstream.

When restoring missing data in the presence of data of the restored indicator and data of analogue points, the input parameters of the ANN for a long period were:

- Data of the restored observation point.
- Data of points-analogs of the restored observation point.

It should be noted that when restoring data, it is possible to use analogue points with data gaps for certain period, but the presence of data for the restored period.

Insufficiency of the volume and quality of the training sample with this restoration method is compensated by the quality of the selection of a suitable analogue item.

In this work, it is proposed to select suitable item(s)-analog(s) using genetic algorithms.
Fig. 4. Scheme for recovering missing data in the presence of data of the recoverable indicator and analogous points for a long-term period.
Adjusting the ANN weights during training and choosing the best ANN model to restore the required indicator is similar to the method of recovering missing data using artificial neural networks in the presence of only data of the restored indicator for a long period.

To restore the hydrological water quality parameters and bottom sediment quality data using this method, the following ANN structures were also used: multilayer perceptron (MLP), radial basis networks (RBF), generalized regression networks (GRNN).

3 Results

Based on the results of evaluating the training of GRNN, RBF, MLP networks (in terms of learning errors, generalization errors, forecast errors on the test sample), a generalized regression neural network (GRNN) was selected to restore gaps in the series of hydrological water quality parameters and bottom sediment quality data. The choice of networks is due to satisfactory learning outcomes (correlation coefficients of calculated and actual values from 0.8 to 0.99), as well as the ability of the GRNN network to consider the cyclicity of values.

For example, the results of training a GRNN network to recover missing hydrological data are presented in Table 1.

**Table 1.** Parameters of the learning outcomes of the GRNN network for the restoration of the Cu content in the water of the Buida river, rep. Bashkortostan, Russia, 2006.

<table>
<thead>
<tr>
<th>Training parameters</th>
<th>Parameter values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network architecture</td>
<td>GRNN 365-338-2-1</td>
</tr>
<tr>
<td>Average absolute difference between real and simulated values</td>
<td>3,85</td>
</tr>
<tr>
<td>Correlation coefficients of calculated and actual values</td>
<td>0,99</td>
</tr>
<tr>
<td>The ratio of the standard deviation of the network error to the standard deviation of the original data</td>
<td>0,11</td>
</tr>
</tbody>
</table>

As you can see from the Table 1, the parameters of the learning outcomes of the GRNN network are satisfactory: the correlation coefficient of real and simulated values is 0.99. This indicates a good convergence of the ANN model of missing data recovery and is the most important indicator of the network for solving this problem.

The trained neural network was used to recover the water quality indicator - sulfates in the Kidysh River downstream of the mouth (Republic of Bashkortostan, Russia) (Figure 5).
Fig. 5. Recovered values of the water quality indicator (sulfates) in the Kidysh river below the mouth (Republic of Bashkortostan, Russia).

4 Discussion

Thus, we can conclude that using traditional methods to recover missing data is not an easy task, since each of them has its own disadvantages. As a result, it is necessary to use artificial intelligence technologies, such as a genetic algorithm (GA), which simulates the processes of natural selection when recovering missing hydroecological data.

The results show that artificial neural networks can serve as one of the adequate tools for restoring gaps in hydrological data, which will significantly improve the quality and speed of information processing, expand their capabilities in applied, research, educational and other tasks related to the analysis and forecast of the ecological state of water bodies.

5 Conclusion

Approbation of the methodology for recovering missing data on water quality and bottom sediment was carried out for the Bayda and Kidysh river in the republic of Bashkortostan, Russia.

Using the Intelligent Problem Solver tool of the Statistica 12.0 program, 5000 ANNs of three architectures with a different number of neurons were built and trained, from which the best models (with the smallest error up to 10%) of each architecture type were selected.

To assess the learning process of the ANN for the restoration of hydrological water quality parameters and bottom sediment quality data, the following parameters were used in the work:

1. Average absolute/relative network error is the average absolute difference between calculated and actual values. If the average relative error is less than 20%, then the network provides good convergence of the calculated and actual values.
2. Correlation coefficient is an indicator characterizing the relationship between real and simulated values. If the value of the coefficient is greater than 0.7, then the network is applicable for forecasting, since it provides good convergence of the calculated and actual values.
3. Network performance (S.D. Ratio) is the ratio of the standard deviation of network errors to the standard deviation of the original data. If the network performance does not
exceed 0.2, then the network is selected well, which is very difficult to achieve due, for example, to the noisy data (inaccurately specified data obtained experimentally).

References


3. Y. Chen et al, Aerosol lidar intercomparison in the framework of the MEMO project. 1. Lidar self calibration and 1st comparison observation calibration based on statistical analysis method, in International Conf. on Meteorology Observations (ICMO), pp. 1-5 (2019)


