Mathematical models of water pollution evaluation

I. Yu. Drozdov1,*, A. V. Aleksakhin2, Yu. V. Aleksakhina3, D. A. Petrusevich1

1Russian Technological University (MIREA), Moscow, 119454, Prospekt Vernadskogo, 78, Russia
2National University of Science and Technology “MISIS”, Moscow, 119049, Leninsky Prospekt, 4, Russia
3Moscow Polytechnic University, Moscow, 107023, Bol’shaya Semenovskaya, 38, Russia

*Corresponding author: drozdovigor94@gmail.com

Abstract. BOD (Biological Oxygen Demand or Biochemical Oxygen Demand) is the amount of dissolved oxygen, expressed e.g. in milligrams of oxygen per dm$^3$ of water, consumed by microorganisms to decompose the organic matter present in water. Ammonium and phosphate contamination prediction in river water and BOD in river water prediction datasets (provided by the Ukrainian government) have been handled. The paper presents proposal of linear regression models (configured with the Ordinary Least Squares (OLS) method) describing dependencies between the target water station and the test station values. The mutual influence between amounts of ammonium and phosphates in river water has been shown The ArDL(p, q) models were used to investigate correlation between concentrations of phosphates and value of BOD at the target river station. Influence of phosphates amount on the BOD in river water has been proven with statistical tests.

1. Introduction

In the modern world ecological problems are more important than ever [1, 2]. With the advent of data science, ecological problems could also benefit from application of data analysis methods. We would like to focus on the problem of water pollution and present a way of employing statistical methods to gain insight into ecological processes.

The Ukrainian Government provides open data on markers of water pollution. This data contains information collected on 8 consecutive water monitoring stations located on the Dnieper river. Provided information includes measurements of three major indicators of water pollution: BOD (Biochemical Oxygen Demand, which is amount of dissolved oxygen that aerobic organisms need to break down organic pollutants), ammonium ions concentrations and phosphate ion concentrations (High phosphate and ammonium levels are usually detected in large cities and agricultural areas and could be sign of sewage contamination). Those measurements are organized as monthly time series data.

We aim to apply statistical and time series analysis methods to investigate spatial and temporal dependencies of pollutant levels. Methods such as ARIMA, ArDL and linear regression are usually applied to economic data [3, 4], but it was also shown that they could be used for forecasting environmental data [5, 6]. In this paper we will use those models to analyze water pollution.
2. Materials and Methods - dataset structure

“The ammonium prediction in river water” (available at https://www.kaggle.com/vbmokin/ammonium-prediction-in-river-water), “The phosphate prediction in river water” (available at https://www.kaggle.com/vbmokin/phosphate-prediction-in-river-water) and “The BOD prediction in river water” (available at https://www.kaggle.com/vbmokin/prediction-bod-in-river-water) datasets have got the same structure. They are based on information about river water collected in Ukraine (by State Water Resources Agency of Ukraine). There are values of concentration at seven water stations and it’s necessary to predict concentration levels at the target station. The closest station to the target one has got number 1. The second one is situated further. Thus, the number denotes distance from the target station. But data of only two first stations have got few omissions. The majority of values at the other stations aren’t measured. So, in the research it’s possible to use only values of concentration of the first two stations and level at the target station that are available in the training set. Values presented in the investigated datasets are average monthly concentrations.

In the dataset there’s a lot of missing values. The original data competition was aimed to get rid of them. The main data series models are constructed with regular data with no omits. So, one of the main problems is to calculate omitted values. It’s done with spline functions. Values of only neighbour measurements by time are used.

Ammonium ions concentration is measured in mg/L. The maximum permissible value of ammonium ions concentration (NH₄) in Ukraine is 0.5 mg/L. Concentration of phosphate ions (polyphosphates) (PₓOᵧ) is measured in mg/L. BOD is measured in mgO₂/L. The maximum permissible value of BOD in Ukraine is 3 mgO₂/L. The statistical measures of parameters in the investigated datasets are shown in the Tables 1, 2, 3.

Table 1. Statistical measures of parameters in the ammonium prediction in river water dataset (in mg/L).

| Parameter | Mean | Standard deviation | Minimum | Maximum |
|-----------|------|--------------------|---------|---------|
| target    | 0.57 | 0.47               | 0       | 2.45    |
| x₁        | 0.56 | 0.49               | 0       | 2.55    |
| x₂        | 0.66 | 0.63               | 0       | 3.60    |

Table 2. Statistical measures of parameters in the phosphate prediction in river water dataset (in mg/L).

| Parameter | Mean | Standard deviation | Minimum | Maximum |
|-----------|------|--------------------|---------|---------|
| target    | 0.31 | 0.29               | 0       | 1.8     |
| x₁        | 0.34 | 0.33               | 0       | 2.39    |
| x₂        | 0.38 | 0.36               | 0       | 1.89    |

Table 3. Statistical measures of parameters in the prediction BOD in river water dataset (in mg/L).

| Parameter | Mean | Standard deviation | Minimum | Maximum |
|-----------|------|--------------------|---------|---------|
| target    | 5.01 | 2.30               | 0       | 11.70   |
| x₁        | 4.99 | 2.21               | 0       | 11.36   |
| x₂        | 5.02 | 2.20               | 0       | 11.00   |

One can notice that concentration mean values increase in case if ammonium and phosphates and decrease in case of BOD. It’s possible to guess that place of pollution is closer to the second station. The course of ammonium concentration time series is presented at the Figure 1. The course of phosphates
concentration time series is presented at the figure 2. The course of BOD time series, its autocorrelation function (ACF) and partial autocorrelation function (PACF) plots [7-10] are presented at the Figures 3-5. Confidence intervals are marked with dashed lines. An ACF shows correlation between time series observations and its values at previous times. A PACF shows the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed. Values of the both functions haven’t got units.

Figure 1. The ammonium ions concentration at the target station time series.

Figure 2. The phosphates concentration at the target station time series.

Figure 3. The BOD level at the target station time series.
3. Results

Linear regression models evaluating BOD and concentration levels at the target station with values at intermediate stations have been constructed and analyzed. Statistical causality tests allow to prove or to refuse dependencies between parameters. The ArDL (autoregressive distributed lag) models are implemented to construct mathematical models describing dependence of the target concentration values and BOD level on concentrations and values of BOD level at \( x_1 \) and \( x_2 \) stations.

3.1. Linear regression models of BOD, ammonium and phosphates concentrations

Linear regressions adjusted with the ordinary least squares method have been constructed in this section. The models are compared with each other by means of the \( R^2 \) determination coefficient value. Also, significance of terms in models should be taken into account \[7, 10\].

The first model tested for each dataset is linear one (the target value is expressed linearly via \( x_1 \) and \( x_2 \) values). Then \( x_1 x_2 \) product, logarithms of variables and their degrees (\( x_1^2, x_2^2 \)) are consequently added to the model if \( R^2 \) grows and inserted terms are significant. Notation \( \hat{t} \) is used for the target variable in expressions. Only the best models are shown in the paper. Expression for ammonium amount at the target station is shown in the equation (1). Of course, the target value of ammonium concentration is expressed with measures of ammonium concentration at two stations. \( R^2 \) value of this model is 81%. Phosphates concentration model is presented in the equation (2), \( R^2 = 53\% \). This model contains only phosphate concentration at different points of the river. Its determination coefficient is very low. There’s no enough data in the dataset to construct better model. BOD model is shown in the equation (3). Its \( R^2 \) is 60%.

\[
\hat{t}_{\text{NH}} = 0.27 + 0.67 x_1 + 0.10 \ln x_2, \quad (1) \\
\hat{t}_{\text{NH}} = -0.51 + 1.87 x_1 - 0.55 x_1^2 - 0.20 \ln x_1, \quad (2) \\
\hat{t}_{\text{BOD}} = 1.72 + 0.47 x_1 + 0.04 x_1 x_2. \quad (3)
\]

Here \( \hat{t}_q \) means evaluation of target time series value, \( x_1 \) and \( x_2 \) mean levels of this value at the intermediate stations.

3.2. ArDL(\( p, q \)) models of dependencies between BOD, ammonium and phosphates concentrations

Time series analysis can also be applied in this task \[9, 10\]. It’s possible to construct ArDL(\( p, q \)) models (autoregressive distributed lag) models describing dependencies between investigated time series. This approach allows to construct linear models similar to the linear regressions but the terms used here are
lags of two time series [11-13]. First of all, existence of such dependence should be proven statistically. The causality tests allowing to understand whether there’s mutual dependence, influence of one time series on the values of another one or there’s no connection at all [14-16].

The Granger test [13-16] has been used to check dependence between BOD value time series, ammonium and phosphates concentration time series. Tests show that there’s mutual dependence between ammonium and phosphates concentrations. The BOD time series depends on the phosphates concentration time series. But there’s no influence of the BOD values on the phosphates concentration series. And at last the BOD time series and the ammonium time series don’t depend on each other.

These results should be considered as intermediate because these time series aren’t full and they still need further analysis. Here one can conclude that growth of water pollution with phosphates causes growth of BOD values showing what quantity of biochemical resources is required to get rid of pollutants. The ArDL models describing dependence of BOD time series on the phosphates time series include the ArDL(2, 2) model (equation (4), $R^2 = 36\%$) and the ArDL(2, 2) model (equation (5), $R^2 = 41\%$) containing $p_{t-2}$ and $p_t$ significant terms (all terms involving BOD values are significant).

\[
BOD_{t} = -0.59 BOD_{t-1} - 0.39 BOD_{t-2} - 1.71 p_t - 0.67 p_{t-1} - 1.51 p_{t-2},
\]

(4)

\[
BOD_{t} = -0.70 BOD_{t-1} - 0.57 BOD_{t-2} - 0.29 BOD_{t-3} - 1.45 p_t - 0.36 p_{t-1} - 1.14 p_{t-2}.
\]

(5)

Here $BOD_{t-q}$ is q-lagged value of BOD time series and $BOD_t$ is its current value; $p_{t-k}$ is k-lagged value of phosphates time series, $p_t$ is its current value. All time series were transformed into stationary form (time difference of 1st order is used). To evaluate these models more thoroughly one needs missing data of various water stations. The BOD time series value tends to decrease. It means that water quality increases after pollution with time.

The same is true for the model of dependence between ammonium and phosphates. There’s mutual dependence between them but because of missing values in the dataset quality of models isn’t high.

Dependence of ammonium concentration on phosphates concentration is shown in the form of the ArDL(2, 2) model in the expression (6). Its $R^2$ is 31%. Dependence of phosphates concentration on ammonium concentration has got the form of the ArDL (2, 2) model and $R^2 = 46\%$.

\[
\hat{t}_{NH_{4}}(\tau) = -0.3 t_{NH_{4}}(\tau - 1) - 0.60 t_{NH_{4}}(\tau - 2) + 0.13 t_{P\cdotO_{2}}(\tau) + 0.32 t_{P\cdotO_{2}}(\tau - 1) + 0.41 t_{P\cdotO_{2}}(\tau - 2),
\]

(6)

\[
\hat{t}_{P\cdotO_{2}}(\tau) = -0.82 t_{P\cdotO_{2}}(\tau - 1) - 0.32 t_{P\cdotO_{2}}(\tau - 2) + 0.06 t_{NH_{4}}(\tau) - 0.04 t_{NH_{4}}(\tau - 1) - 0.13 t_{NH_{4}}(\tau - 2).
\]

(7)

Here $\hat{t}_{NH_{4}}(\tau)$ means evaluation of target time series evaluating ammonium concentration value, $t_{NH_{4}}(\tau - i)$ is its i-lagged value. $\hat{t}_{P\cdotO_{2}}(\tau)$ is evaluation of target time series evaluating phosphates concentration value, $t_{P\cdotO_{2}}(\tau - i)$ is its i-lagged value, $\tau$ denotes dependence from time.

There are negative signs in the first two terms in the equations (6) and (7). It can be considered as decrease of pollutant concentration with time.

4. Conclusion

The investigated datasets contain information about ammonium ions concentration, phosphate ions concentration in river water and evaluation of BOD in river water. Analysis of dependence shows that BOD is mostly influenced by phosphate concentration, while no significant impact of ammonium concentration on BOD was discovered. At the same time there exists mutual dependence between ammonium ions concentration and phosphates concentration in river water, which may indicate compound pollution. Thus, pairs of these time series change together. This result confirms correctness of the models though it’s based only on the data of two intermediate water stations and the target one from overall eight consequent water stations.
Linear regression models of BOD level, ammonium and phosphates amounts in river water have been constructed. The models (1) – (3) describe dependence of values at the target station on values at the intermediate ones. BOD target level is evaluated with levels of BOD at the intermediate stations. The same is true for the other models. Determination coefficients of these models are between 53% and 81%. They would be higher if the dataset was full. Coefficients of the ammonium estimation model (1) are positive. It means that limit of ammonium concentration is exceeded often.

The ArDL models describing connection between the BOD value and phosphates concentration level have been constructed (expressions (4), (5)). BOD value decreases with time. It can be considered as steady recycling of water. Also, mutual dependence between ammonium concentration time series behaviour and phosphates concentration time series values has been proved with statistical tests. These models show that values of concentration tend to decrease with time. Water becomes cleaner with time. Properties of these models would be better if the dataset were full and concentrations at all water stations were presented.

At the same time constructed mathematical models work even if there’s lack of data from five water stations. They work with use of only target station data and data from two water stations. Thus, proposed technique works even in the conditions of missing data from a few water stations. One is able to prove statistical dependence between concentrations and BOD level time series and to construct mathematical models of their mutual dependencies of “cause-and-effect” dependence with certain level of quality.

Construction of models based on information with missing values is very important in a lot of domains of knowledge. Lack of data can be caused with broken equipment, wrong measuring methodology. Some part of data can be removed if changes made by local authorities are suspected.

References
[1] Petrusevich D 2019 SGEM 2019 19(2.1) pp 673-680
[2] Wen M, Li P, Zhang L, Chen Y 2019 IEEE Access PP(99) 1
[3] Sidorova E 2019 SGEM 2019 19(5.3) pp 531-538
[4] Eliseeva E 2019 SGEM 2019 19(5.3) pp 307-316
[5] Abhilash M, Thakur A, Gupta D, Sreevidya B 2018 Advances in Intelligent Systems and Computing 696 pp 413-426
[6] Rahman M A and Hossain M A 2012 Asian J of Water, Environment and Pollution 9 pp 63-69
[7] James G, Witten D, Hastie T, Tibshirani R 2015 An introduction to statistical learning with applications in R Springer-Verlag New York 426 p
[8] Sidorova E 2019 SGEM 2019 19(5.3) pp 841-848
[9] Shrestha M B and Bhatta G R 2018 The J. of Finance and Data Sci 4(2) pp 71-89
[10] Hyndman R J and Athanasopoulos G 2018 Forecasting: principles and practice Otexts 2 edition 382 p
[11] Kalugin T R, Kim A K, Petrusevich D A 2020 Russ Tech J 8(2) pp 7-22 [in Russ]
[12] Nkoro E and Uko A K 2016 J. of Stat and Econometric Methods 5(4) pp 63-91
[13] Pesaran M H and Shin Y 1999 An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis Cambridge University Press Cambridge
[14] Granger C W J 2001 Essays in Econometrics: The Collected Papers of Clive W.J. Granger Cambridge University Press
[15] Hatemi J A 2008 Empirical Economics 35(3) pp 497-505
[16] Engle R and Granger C 1991 Long-Run Economic Relationships: Readings in Cointegration Oxford University Press