Analysis of ammonium and phosphates concentration time series in river water

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. The Ammonium prediction in river water and the Phosphate prediction in river water datasets (provided by the Ukrainian government) have been handled. The ARIMA (p, d, q) model of low order (p<5 and q<5) and high-order models (p>4 or q>4) have been constructed to describe behaviour of concentration time series at different river stations. Orders and coefficients in the first case are obtained via information criteria minimization. In the second case the ACF and PACF functions analysis is used. The ArDL(p, q) models are used to investigate correlation between concentrations of ammonium and phosphates at different river stations. Dependence of these parameters at the target water station and at other stations is confirmed with statistical tests. Conclusions about water pollution are made.

1. Introduction
In the modern world a lot of concern is raised over ecological problems: forest fires and deforestation, global warming, water pollution etc. [1, 2]. With the advent of advanced data science and time series analysis techniques, they provide key instruments for description and forecasting of various processes. Statistical methods and time series analysis are often implemented to investigate parameters of water streamflow and to predict water consumption [3, 4]. ARIMA and ArDL time series models are usually used in analysis of physical and economic data but nowadays they are also applied to describe environmental data in various domains. In particular, ARIMA (p, d, q) models are widely used in order to investigate time series behaviour and to make forecasts. This technique is implemented in various domains of knowledge and it’s successfully used to describe processes in mining industry and predict metal price changing [9, 10], behaviour of river water level or chemical substances in water levels and to predict further changes of such values [3-5] and so on.

In the paper we investigate river water pollution by applying ARIMA and ArDL models to water monitoring datasets. Used technique can be implemented in description of other ecological processes and problems.
2. The dataset structure

In this paper analysis of time series is implemented in order to investigate pollution of river water. Mathematical models based on statistics are able to detect pollution and in some cases it’s possible to predict future levels of pollutants concentrations. The investigated datasets contain information about ammonium ions concentration [6] and phosphate ions concentration [7] in river water. The geographic information systems [8] play a crucial role in gathering reliable data for environmental investigations and decision making process. Values of concentration are measured at eight consequent water stations. The original data contest was aimed to predict concentration level at one of the stations using values at the other ones. In the paper mathematical models describing behaviour of concentration level time series at the target station are built. Also, models aimed to describe dependence of concentration at the target station on levels at the other stations are constructed and investigated. Information from consequent water stations isn’t full. First of all, measurements from only three stations are available (they are denoted as \(x_1\) and \(x_2\) intermediate stations and the target one). Data from other stations are almost completely unavailable. So, in this research all models use data of three stations. At the stage of dataset preparation some values were removed manually. In the original data competition one of the main problems is to handle missing values. This problem is also solved in the research. The cubic spline functions are used to get rid of missing values. The proposed technique can be used to investigate pollution levels at plants in other domains of industry.

The ammonium [6] and the phosphate [7] predictions in river water datasets are of the same structure. These datasets include information on river water collected in Ukraine (by State Water Resources Agency of Ukraine). There are values of concentrations at seven water stations and it’s required to predict those values at the target station. The closest station to the target one has got number 1. The second one is situated further upstream. Thus, the number denotes distance from the target station. Data of only two first stations is nearly complete. 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 is available in the train set. The values denote average monthly concentrations. The concentration of ammonium ions is measured in milligrams per cubic decimeter. The maximum permissible value of ammonium ions concentration (NH4) in Ukraine is 0.5 mg/L [6]. The concentration of phosphate ions (polyphosphates) (PxOy) is measured in mg/L [7]. The statistical measures of parameters in the datasets [6, 7] are shown in Tables 1, 2.

| 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    |
| \(x_1\)  | 0.56 | 0.49               | 0       | 2.55    |
| \(x_2\)  | 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     |
| \(x_1\)  | 0.34 | 0.33               | 0       | 2.39    |
| \(x_2\)  | 0.38 | 0.36               | 0       | 1.89    |
3. Experiments

First of all, the target, $x_1$ and $x_2$ time series contacting concentration levels at the target station and at the first two water stations respectively are analyzed. The ARIMA (p, d, q) (autoregressive integrated moving average) models are used to describe these time series behaviour. The ArDL (autoregressive distributed lag) models are implemented to construct mathematical models describing dependence of the target concentration value on concentrations at $x_1$ and $x_2$ stations.

3.1. Ammonium concentration prediction with ARIMA (p, d, q) models

The ARIMA mathematical model construction is usually based on the Box-Jenkins method [11-13]. The behaviour of the target time series is presented in Figures 1-3. First of all, it must be transformed into a stationary form. Stationarity of the 1st-order difference is confirmed with the KPSS test [11, 12]. For this series number of difference is equal to 1 [11-14].

The same technique is used to handle $x_1$ and $x_2$ time series. They are both stationary. Seasonality of the time series [11, 12] is usually analyzed with autocorrelation function (ACF) and partial autocorrelation function (PACF) plots shown in Figures 2, 3 (confidence intervals are marked with dashed lines). Though the time series is monthly there’s no obvious monthly or yearly seasonality.

In the next step of time series analysis the $p$, $q$ orders need to be obtained. The most widely used technique supposes implementation of the Akaike information criterion and Bayes information criterion [11, 12]. The best fitted models should have the lowest values of criteria. At the same time the orders are usually confirmed with ACF and PACF plots analysis [11, 12]. According to this approach the $x_1$ has got the ARIMA (1, 1, 2) model and $x_2$ time series is ARIMA (0, 1, 1) simple exponential smoothing [12] the target one can be explained with the ARIMA (0, 1, 1) random walk model. It’s presented in the Table 2.

One of other approaches is based only on the ACF and PACF plots analysis [15]. The information criteria used for order selection also include penalty for high-order models [12, 15].

![Figure 1. The ammonium ions concentration at the target station time series.](image)

![Figure 2. The ACF function of the target time series.](image)

![Figure 3. The PACF function of the target time series.](image)
If this penalty is removed, the high-order ARIMA models \((p>4 \text{ and } q>4)\) can be constructed to describe time series behaviour. The data has been divided into two parts. The first one has been used as a training set to construct ARIMA models. The second one contains the last three values of the time series and it’s used as a test set. The RMSE (root mean square error) and MAE (mean average error) are counted at the test set and mean value for the whole test period is obtained. The best target ARIMA model obtained with this approach is presented in the Table 3. Because of the maxima at the 7th lag at the ACF function plot the models of \(p = 7 \text{ and } q = 7\) orders and lower have been tested. The best model is ARIMA \((7, 1, 7)\). Though information criteria values are higher for this model, the RMSE value of the forecast is better and the MAE value is approximately at the same level as for the ARIMA \((0, 1, 1)\) model.

### Table 3. The ARIMA \((p, d, q)\) models of ammonium concentration levels at the target water station.

| ARIMA model | Akaike criterion value | Bayes criterion value | RMSE (error of prediction) | MAE (error of prediction) |
|-------------|------------------------|-----------------------|-----------------------------|--------------------------|
| ARIMA(0, 1, 1) | 152.13 | 158.06 | 1.04 | 0.30 |
| ARIMA(7, 1, 7) | 163.68 | 208.13 | 0.49 | 0.39 |

### 3.2. Phosphates concentration prediction with ARIMA \((p, d, q)\) models

The same technique has been used to construct phosphates concentration mathematical models. The target time series containing values of phosphates concentration at the target station and its ACF and PACF plots are presented in Figures 4-6. According approach based on information criteria implementation, the \(x_1\) and \(x_2\) time series are ARIMA \((0, 1, 1)\) models (simple exponential smoothing) [12] and the target time series has got the ARIMA \((1, 1, 1)\) form. Stationary property of the 1st-order differenced time series has been confirmed with the KPSS test [11, 12].

![Figure 4. The phosphates concentration at the target station time series.](image1)

![Figure 5. The ACF function of the target time series.](image2)

![Figure 6. The PACF function of the target time series.](image3)
The analysis of ACF and PACF plots shows that possible values of $p, q$ orders are limited with 7. Thus, the ARIMA models of orders $p \leq 7$ and $q \leq 7$ have been tested. They have been compared with quality of forecast for the period of 3 months. The best model has got the ARIMA (1, 1, 4) form. Models constructed with use of various approaches are presented in the Table 4. Though information criteria values are higher the RMSE value of the forecast is better and the MAE value is approximately at the same level as for the ARIMA (1, 1, 1) model.

| ARIMA model | Akaike criterion value | Bayes criterion value | RMSE (error of prediction) | MAE (error of prediction) |
|-------------|------------------------|-----------------------|----------------------------|--------------------------|
| ARIMA(1, 1, 1) | 32.21 | 41.1 | 0.68 | 0.44 |
| ARIMA(1, 1, 4) | 37.55 | 55.32 | 0.42 | 0.41 |

### 3.3. ArDL (p, q) models of dependencies between time series

In the previous sections dependence of time series value on its previous values has been investigated. The ARDL(p, q) model (autoregressive distributed lag) is a tool describing dependence between two time series [15-18]. First of all, such dependence should be confirmed with statistical tests. Here the Granger test is used [15, 18-22]. The ARDL(p, q) models have been constructed for the ammonium concentration dataset [6]. The 1st-order differences time series of target, $x_1$ and $x_2$ have been handled with cointegration tests. The statistical tests show that there’s no causal connection between these variables. This result needs some explanation and it’s shown further in the phosphates concentration models analysis. Still the ARDL models have been tested and the best one describes possible connection between target and $x_1$ time series. It has got the form of ARDL(3, 2) model (the ARDL(3, 3) model has been tested but the highest order term has been insignificant, and it has been removed from the model). The $R^2$ value of this model is 82%. It’s shown at the equation (1):

$$
\hat{\text{NH}_4} (\tau) = -0.65 \text{NH}_4 (\tau - 1) - 0.24 \text{NH}_4 (\tau - 2) + 0.84 x_2 (\tau) + 0.62 x_2 (\tau - 1) + 0.28 x_2 (\tau - 2) + 0.11 x_2 (\tau - 3).
$$

Here $\text{NH}_4 (\tau)$ means evaluation of ammonium ions concentration target time series value, $\text{NH}_4 (\tau - i)$ is its i-lagged value, $x_2 (\tau)$ is current value $x_2$ at the second station, $x_i (\tau - i)$ is its i-lagged value. The ARDL(p, q) models have been constructed for the phosphates concentration dataset [7]. The 1st-order differences time series of target, $x_1$ and $x_2$ have been handled with cointegration tests. It’s curiously that dependence between target and $x_1$ station values hasn’t been confirmed. There’s also no statistical dependence between $x_1$ and $x_2$ station values. But at the same time dependence between target and $x_2$ station values has been proven with this test.

In our opinion these results can be explained in two ways. First of all, there’s a lot of omitted data at other stations. So, they haven’t been included into this research. There also can be source of pollution between the target station and the first one $x_1$. The influence of this possible pollution decreases with growth of distance. That’s why statistical dependence between target and $x_1$ still exists.

Models evaluating phosphates concentration has also been investigated. The statistical tests show mutual dependence between target and $x_2$ station values. Also, target station time series values depend on $x_2$ station values. Thus, expected dependencies exist though there are omitted data. The ARDL(p, q) models of the $p = 3$, $q = 3$ and $p = 4$, $q = 4$ orders have got the best parameters. Models of higher orders have got insignificant terms of highest order. Models of lower orders have got less $R^2$ determination coefficient values. The ARDL(3, 3) model has got $R^2 = 65\%$ and it’s presented in the equation (2). The ARDL(4, 4) one has got value of $R^2 = 67\%$ and it’s shown in the equation (3).

$$
\hat{\text{PO}_4} (\tau) = -0.76 \text{PO}_4 (\tau - 1) - 0.48 \text{PO}_4 (\tau - 2) - 0.18 \text{PO}_4 (\tau - 3) + 0.48 x_2 (\tau) + 0.26 x_2 (\tau - 1) + 0.35 x_2 (\tau - 2) + 0.21 x_2 (\tau - 3),
$$

### Table 4. The ARIMA (p, d, q) models of phosphates concentration levels at the target water station.
\[ \hat{t}_{P_{O_2}}(\tau) = -0.81 t_{P_{O_2}}(\tau - 1) - 0.62 t_{P_{O_2}}(\tau - 2) - 0.33 t_{P_{O_2}}(\tau - 3) - 0.23 t_{P_{O_2}}(\tau - 4) + 0.52 x_2(\tau) + 0.35 x_2(\tau - 1) + 0.46 x_2(\tau - 2) + 0.33 x_2(\tau - 3) + 0.21 x_2(\tau - 4). \] (3)

Here \( \hat{t}_{P_{O_2}}(\tau) \) means evaluation of phosphates concentration target time series value, \( \hat{t}_{P_{O_2}}(\tau - i) \) is its i-lagged value, \( x_2(\tau) \) is current value \( x_2 \) at the second station, \( x_2(\tau - i) \) is its i-lagged value. One can assume that these models could have better parameters if the dataset was full.

4. Conclusion

The ammonium and phosphates prediction in river water [6, 7] datasets allow to evaluate water pollution levels in river water in Ukraine. The statistical values of the dataset [6] shown in the Table 1 confirm that there’s pollution. There is concentration of ammonium ions above normal values. The mathematical models of target concentration value have been built. The ARIMA models present dependence on previous values of the investigated time series. Even though data of stations 3-7 is mostly absent, data of stations 1 and 2 is suitable for explaining and forecasting target series. Missing values were interpolated using locally fitted splines.

The ArDL models express dependence between the target time series and pollution levels at other river water stations. One can notice that the datasets aren’t full. They contain information only from two stations and from the target one. Thus, the best ArDL models connecting these values have got R² value about 67% for the phosphates concentration estimation and 82% for ammonium concentration estimation. Because of omitted data the cointegration test shows statistical dependence between target station values and other station values only in the case of phosphates concentration investigation. At the same time this point is another indirect prove of pollution that can take place between the stations.

A lot of attention is paid to the environment and water pollution problems nowadays. Sustainable activities [23] demand for reliable data for optimal choice of remediation technology. This research shows once more that statistical approach can highlight them in the process of open data analysis.

References

[1] Sidorova E 2019 SGEM 19(5.3) pp 531-38
[2] Eliseeva E 2019 SGEM 19(5.3) pp 307-16
[3] Razali S N A M, Rusiman M S, Zawawi N I, Arbin N 2018 J. of Physics: Conf Series 995 012041
[4] Najah N, Samsudin R, Shabri A 2017 J. of Physics: Conf Series 890 012141
[5] Abhilash M, Thakur A, Gupta D, Sreevidya B 2018 Advances in Intelligent Systems and Computing 696 pp 413-26
[6] Ammonium prediction in river water Retrieved from (the 1st of September, 2020): https://www.kaggle.com/vbmokin/ammonium-prediction-in-river-water
[7] Phosphate prediction in river water Retrieved from (the 1st of September, 2020): https://www.kaggle.com/vbmokin/phosphate-prediction-in-river-water
[8] Kulikova E and Ivanikov A 2019 13th Int. Sci. Conf. Monitoring 160684
[9] Kriechbaumer T, Angus A, Parsons D, Casado M R 2014 Resources Policy 39 pp 32-41
[10] Sarkisov S S, Lomonosova N V, Zolkina A V, Sarkisov T S 2020 Tsvetnye Metally 3 (in Russ)
[11] Hyndman R J and Khandakar Y 2008 J. of Stat Software 27(3) pp 1-22
[12] Hyndman R J and Athanasopoulos G 2018 Forecasting: principles and practice Otexts 2 edition 382 p
[13] Box G and Jenkins G 2008 Time series analysis: Forecasting and Control John Wiley and Sons 712 p
[14] Said S E and Dickey D A 1984 Biometrika 71(3) pp 599-607
[15] Petrucevich D 2019 SGEM 2019 19(2.1) pp 673-80
[16] Kalugin T R, Kim A K, Petrucevich D A 2020 Russ Tech J 8(2) pp 7-22 (In Russ)
[17] Nkoro E and Uko A K 2016 J. of Stat and Econometric Methods 5(4) pp 63-91
[18] Pesaran M H and Shin Y 1999 *An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis* Cambridge University Press Cambridge

[19] Granger C W J 1980 *J. of Economic Dyn and Control* 2 pp 329-52

[20] Granger C W J 2001 *Essays in Econometrics: The Collected Papers of Clive W.J. Granger* Cambridge Cambridge University Press

[21] Hatemi J A 2008 *Empirical Economics* 35(3) pp 497-505

[22] Engle R and Granger C 1991 *Long-Run Economic Relationships: Readings in Cointegration* New York Oxford University Press

[23] Ivannikov A L, Kongar-Syuryun C, Rybak J, Tyulyaeva Yu 2019 *IOP Conf. Ser.: Earth Environ. Sci.* 362 012130