Water pollution time series analysis

I Yu Drozdov and D A Petrusevich
MIREA – Russian Technological University, Prospekt Vernadskogo, 78, Moscow, 119454, Russia

E-mail: petrdenis@mail.ru

Abstract. The time series of the ammonium ions, phosphates in river water and the BOD (biochemical oxygen demand) prediction in river water are analyzed in the paper. These time series have got missing values. So, the first problem is to handle missing data. Splines are implemented to fill the data. After that the mathematical models of each time series are tested and models of dependencies between time series are analyzed. The ARIMA (p, d, q) models are used to explain time series behavior with its previous values. The ArDL (p, q) models and linear regression models are constructed in order to describe dependencies between time series. Quality of river water is very important ecological problem. Mathematical models are able to provide analysis of water pollution. The proposed methods can be used in other problems of analysis of time series with missing data.

1. Introduction
Ecology and environmental sciences problems are very important in modern world. Time series analysis methods, particularly an ARIMA family of models and related Box-Jenkins methodology are powerful methods developed in econometrics field to describe complex time-dependent behavior of different processes [1 – 3]. Also, it has been shown that such models could be used for analysis and prediction of ecological data [4]. In particular, ARIMA models have been successfully applied to air pollution data around the world [5 – 7]. Aim of this paper is to apply time series analysis methods to investigate water pollution in Ukraine. We are using datasets provided by the Ukrainian government (State Water Resources Agency of Ukraine) which contain information about biochemical oxygen demand (BOD) [8], ammonium ions concentration [9] and phosphate ions concentration [10] in river water. 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. Time series values have got periods of missing data. First of all, these missing parts need to be calculated. Spline functions have been used to solve this task. In the paper mathematical models describing behaviour of concentration and BOD 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. Influence of pollution values at intermediate stations on levels at the target stations are investigated with statistical tests. The proposed technique can be used to investigate pollution at plants in other domains of industry and to handle missing values in various time series.
2. The dataset structure

We are concerned with 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). We will refer to these datasets as BOD, ammonium and phosphate. Presented datasets are of the same structure. There are values of seven water station used to predict pollution level at the eighth one. 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. But there are two problems. Data of only two first stations is nearly complete. Values of other stations aren’t available. Also there are missing values in the data of the target and two intermediate stations. Missing values should be filled. After that it’s possible to investigate dependence between time series. Values denote average monthly concentrations. Ammonium ions concentration is measured in milligrams per cubic decimeter. The maximum permissible value of ammonium ions concentration (NH4) in Ukraine is 0.5 mg/cub. dm [9]. Concentration of phosphate ions (polyphosphates) (PxOy) is measured in mg/cub. dm [10]. The maximum permissible value of BOD in Ukraine is 3 mgO/cub. dm [8].

The missing values in all time series have been handled by means of spline functions. Splines have been constructed using known time series values closest to the missing ones. The statistical measures of parameters in the datasets [8–10] are calculated after the missing values are handles and they are shown in the tables 1, 2, 3.

The figures 1 – 3 demonstrate behaviour of the ammonium ions concentration [9] and the table 1 demonstrates its statistical parameters. The figures 4 – 6 demonstrate behaviour of the phosphate ions concentration [10] and the table 2 demonstrates its statistical parameters. The figures 7 – 9 demonstrate behaviour of the BOD time series [9] and the table 3 demonstrates its statistical parameters.

| 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    |

| 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    |

| 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   |

Maximum values in these tables show that there’s pollution above limits. In case of ammonium and phosphates concentrations evaluation it’s seen that mean values and standard deviation at the second station x2 are higher. This fact can point at pollution problems in this area.
3. Experiments
After missing values are handled it’s possible to construct mathematical models aiming to evaluate time series and to investigate dependencies between them.

3.1. ARIMA (p, d, q) models of the investigated time series
All time series handled in this research are presented in the form of ARIMA models. These results are shown in the tables 4 – 6. Levels of information criteria measure should be less for the models fitted to the investigated time series values better. At the same time they minimize orders p, q of the ARIMA(p, d, q) models. This question is observed in [2, 11].

Table 4. The ARIMA (p, d, q) models of the Ammonium prediction in river water dataset time series.

| Parameter | ARIMA model | AIC (Akaike information criterion value) | BIC (Bayes information criterion value) |
|-----------|-------------|------------------------------------------|------------------------------------------|
| target    | ARIMA(5, 1, 1) | 199.63 | 223.19 |
| $x_1$     | ARIMA(0, 1, 2) | 279.53 | 289.63 |
| $x_2$     | ARIMA(1, 0, 4) | 393.89 | 417.48 |

Table 5. The ARIMA (p, d, q) models of the Phosphate prediction in river water dataset time series.

| Parameter | ARIMA model | AIC (Akaike information criterion value) | BIC (Bayes information criterion value) |
|-----------|-------------|------------------------------------------|------------------------------------------|
| target    | ARIMA(0, 1, 3) | 6.95 | 20.41 |
| $x_1$     | ARIMA(1, 1, 2) | 100.23 | 113.69 |
| $x_2$     | ARIMA(2, 1, 3) | 121.69 | 141.88 |

Table 6. The ARIMA (p, d, q) models of the Prediction BOD in river water dataset time series.

| Parameter | ARIMA model | AIC (Akaike information criterion value) | BIC (Bayes information criterion value) |
|-----------|-------------|------------------------------------------|------------------------------------------|
| target    | ARIMA(2, 0, 1) | 909.42 | 926.27 |
| $x_1$     | ARIMA(0, 0, 1) | 970.18 | 980.29 |
| $x_2$     | ARIMA(0, 0, 1) | 945.87 | 955.98 |

In case of Prediction BOD dataset time series of the intermediate stations values are just MA (1) processes. It makes forecasting of their values impossible. The same is true for the ARIMA (0, 1, 3) model of the phosphates target station time series. Two last groups of time series have got the same degree of differencing.

Figure 1. The ACF function of the ammonium target time series.
Ammonium series is non-stationary, so order 1 differencing is required to achieve stationarity. ACF is trailing off, and PACF function cuts off. This indicates mostly autoregressive model, which is confirmed by information criterion fitting procedure.
Plot of phosphate concentration shows that actual levels are regularly exceeding maximum permissible value of 0.5 mg/cub. dm. and therefore water pollution is present. Notable observation of ACF and PACF is that there are spikes in correlations at regular intervals of 3 months. This suggests that quarter seasonality might be present. We were able to fit ARIMA models without seasonality before because first-order differencing was able to make the series stationary, but we could have lost important information. We have fitted new SARIMA models with assumed quarter seasonality. Results are shown in table 7.

Table 7. The SARIMA (p, d, q) (P, D, Q) models of the Phosphate prediction in river water dataset time series.

| Parameter | SARIMA model | AIC (Akaike information criterion value) | BIC (Bayes information criterion value) |
|-----------|--------------|------------------------------------------|-----------------------------------------|
| target    | ARIMA(0,1,2)(0,0,2)[3] | 5.12                                     | 21.95                                   |
| $x_1$     | ARIMA(1, 1, 2) | 100.23                                   | 113.69                                  |
| $x_2$     | ARIMA(1,1,2)(2,0,1)[3] | 122.57                                   | 146.13                                  |

Most notable improvement of seasonality introduction demonstrated by the model of phosphate ions concentration target series. Series $x_1$ is still best described by non-seasonal model.

![Figure 7. The ACF function of the BOD target time series.](image1)

![Figure 8. The PACF function of the BOD target time series.](image2)

![Figure 9. The BOD ions concentration at the target station time series.](image3)

3.2. Dependencies between the time series
To describe dependencies between time series, first of all, it’s necessary to prove cause-and-effect relationship. The Granger causality tests have been implemented to all time series in [1 – 3] datasets. It
can show whether changes in values of one time series can imply changes in other time series behaviour, if there’s mutual dependence between them or they behave independently from each other [12 – 16].

The Granger test has confirmed mutual dependence between BOD target and phosphates target time series. There are no dependencies between other pairs of time series including values at the target water station.

The time series measuring values of BOD, ammonium and phosphates concentration have been also checked separately.

The BOD target time series depends on the BOD \(x_1\) station time series.

There’s mutual dependence between the ammonium target time series and intermediate station time series. But values at the \(x_1\) and \(x_2\) stations are independent.

The \(\text{ArDL}(p, q)\) models constructing dependence behaviour between ammonium target time series and its values at the intermediate stations have been tested [17, 18]. The orders \(p\) and \(q\) have been increased until the highest order terms have become insignificant. After that stage the highest terms have been deleted from the model until the highest order term become significant again. The \(\text{ArDL}(2, 2)\) model structure connecting ammonium target time series with values at the intermediate stations is shown in the equation (1):

\[
\hat{q}_t = 0.07 + 0.57q_{t-1} - 0.19q_{t-2} + 0.49x_{p_1} - 0.10x_{p-1} + 0.12x_{p-2}. \tag{1}
\]

Here \(\hat{q}_t\) means evaluation of ammonium target time series value, \(q_{t-i}\) is its \(i\)-lagged value, \(x_{p_1}\) is current value \(x_1\) at the first station, \(x_{p-1}\) is its \(i\)-lagged value. \(R^2\) value of this model is 62%.

The same procedure has been made for other dataset about phosphates concentration and prediction of BOD.

Investigation of the phosphates concentration dataset leads to the \(\text{ArDL}(2, 3)\) model. One can see its structure in the expression (2):

\[
\hat{q}_t = 0.04 + 0.33q_{t-1} - 0.25q_{t-2} + 0.14q_{t-3} + 0.39x_{p_1} - 0.08x_{p-1} + 0.23x_{p-2}. \tag{2}
\]

Here \(\hat{q}_t\) means evaluation of phosphates target time series value, \(q_{t-i}\) is its \(i\)-lagged value, \(x_{p_1}\) is current value \(x_2\) at the second station, \(x_{p-1}\) is its \(i\)-lagged value. \(R^2\) value of this model is 51%. All terms in (2) are significant except the \(x_{p-1}\) term.

The best \(\text{ArDL}(p, q)\) model for the BOD prediction dataset is the \(\text{ArDL}(2, 3)\) model:

\[
\hat{q}_t = 1.93 + 0.54q_{t-1} - 0.25q_{t-2} - 0.11q_{t-3} + 0.44x_{p_1} - 0.15x_{p-1} + 0.14x_{p-2}. \tag{3}
\]

Here \(\hat{q}_t\) means evaluation of the BOD prediction target time series value, \(q_{t-i}\) is its \(i\)-lagged value, \(x_{p_1}\) is current value \(x_1\) at the first station, \(x_{p-1}\) is its \(i\)-lagged value. \(R^2\) value of this model is 47%. All terms in (3) are significant.

Low values of determination coefficient in equations (1) – (3) show that mathematical models predictions values at the target stations should include information from all seven concurrent water stations. In these experiments only information from two of them is used because in case of other ones there are almost no measured values at all.

Causal connections between time series of ammonium, phosphates ions concentration and BOD prediction target station values. The BOD target station level depends on the phosphates ions concentration. The \(\text{ArDL}(2, 3)\) model of this connection is shown in the expression (4):

\[
\text{BOD}_t = 4.09 + 0.59\text{BOD}_{t-1} - 0.25\text{BOD}_{t-2} - 0.15\text{BOD}_{t-3} + 1.05p_1 - 0.96p_{t-1} + 1.49p_{t-2}. \tag{4}
\]
Here $BOD_{t,q}$ is q-lagged value of BOD time series and $BOD_t$ is its current value; $p_{k,t}$ is k-lagged value of phosphates time series, $p_t$ is its current value. $R^2$ value of this model is 33%. All terms in (4) are significant.

Linear regressions adjusted with the ordinary least squares method have been constructed in this section. The models are compared by value of $R^2$ determination coefficient. Significance of terms is taken into account [19, 20].

The first model tested for each dataset is linear one (the target value is expressed linearly via $x_1$ and $x_2$ values). Then $x_1x_2$ product, logarithms of variables and their degrees ($x_1^2$, $x_2^3$) are consequently added to the model if $R^2$ grows and inserted terms are significant. Notation $t$ is used for the target variable in expressions. Though there are a few models with $R^2$ value close to the maximal level only the best models are shown in the paper.

The models of ammonium ions concentration at the target station are shown in the equations (5), (6). The first one has got $R^2 = 52\%$. $R^2$ of the second model is 53%.

$$\hat{i}_a = 0.05 + 0.70x_1 + 0.41x_2 - 0.23x_1x_2,$$  

(5)

$$\hat{i}_a = 0.07 + 1.62x_1 - 2.15x_1^2 + 1.83x_1^3 - 0.47x_1^4.$$  

(6)

Here $\hat{i}_a$ means evaluation of target time series value of ammonium ions concentration, $x_1$ and $x_2$ mean levels of this value at the intermediate stations.

Models of the same structure have been tested for the phosphates ions concentration dataset. The best model is shown in the expression (7). Its $R^2$ is 45%.

$$\hat{i}_p = 0.23 - 0.95x_1 + 3.36x_1^2 - 2.79x_1^3 + 0.66x_1^4 + 0.55x_1^2 - 0.25x_1^3.$$  

(7)

Here $\hat{i}_p$ means evaluation of target time series value of phosphates ions concentration, $x_1$ and $x_2$ mean levels of this value at the intermediate stations.

Linear regression models constructed to describe the BOD level have got determination coefficient value less than 30%.

4. Conclusion

The investigated datasets contain information about ammonium ions concentration [9], phosphate ions concentration [10] in river water and evaluation of BOD [8] in river water. There are missing values in the datasets. They have been handled with spline functions. Statistical tests show dependence between these values. Dependence of BOD value on the phosphates concentration has been proved. Regular excess of maximum permissible values and therefore water pollution is present.

ARIMA models describing time series behavior have been fitted. Phosphate ions concentration demonstrate quarter seasonality, which could indicate correlation with industrial processes cycles.

Linear regression models of ammonium and phosphates ions concentration level in river water have been constructed. The models (5) – (7) describe dependence of values at the target station on values at the intermediate ones. Determination coefficients of these models are between 45% and 53%. Low determination coefficients in regression model between target station and stations 1 and 2 could indicate that pollution is introduced somewhere on this interval.

The ArDL models describing connection between the BOD value and phosphates concentration level have been constructed (expression (4)). Also, mutual dependence between ammonium concentration time series behaviour and phosphates concentration time series values has been proved with statistical tests.

The ArDL models of dependencies between values at the intermediate stations and at the target one for ammonium, phosphates and BOD values are shown in the equations (1) – (3).
References

[1] Andrianova E G, Golovin S A, Zykov S V, Lesko S A and Chukalina E R 2020 Review of modern models and methods of analysis of time series of dynamics of processes in social, economic and socio-technical systems *Russ. Technological J.* 8(4) 7-45 doi: 10.32362/2500-316X-2020-8-4-7-45

[2] Petruzevich D A 2019 Analysis of mathematical models used for econometrical time series forecasting *Russ. Technological J.* 7(2) 61-73 doi: 10.32362/2500-316X-2019-7-2-61-73

[3] Sarkisov S S, Lomonosova N V, Zolkina A V and Sarkisov T S 2020 Integration of digital technology in mining and metallurgy industries *Tsvetnye Metally* 3 7-14 doi: 10.17580/tsm.2020.03.01

[4] Ives A, Abbott K and Ziebarth N 2010 Analysis of ecological time series with ARMA (p, q) models *Ecology* 91 858-71 doi: 10.1890/09-0442.1

[5] Abhilash M, Thakur A, Gupta D and Sreevidya B 2018 Time Series Analysis of Air Pollution in Bengaluru Using ARIMA Model *Ambient Communications and Computer Systems. Advances in Intelligent Systems and Computing* 696 413-26 10.1007/978-981-10-7386-1_36.

[6] Rahman M A and Hossain M A 2012 Time series analysis model for particulate matter of air pollution data in Dhaka city *Asian J of Water, Environment and Pollution* 9 63-9

[7] Guarnaccia C, Griselda C-B, Breton R, Tepedino C, Quartieri J and Mastorakis N 2018 ARIMA models application to air pollution data in Monterrey, Mexico *AIP Conf Proc* 1982 p 020041 doi: 10.1063/1.5045447.

[8] Prediction BOD in river water 2020 https://www.kaggle.com/vbmokin/prediction-bod-in-river-water

[9] Ammonium prediction in river water 2020 https://www.kaggle.com/vbmokin/ammonium-prediction-in-river-water

[10] Phosphate prediction in river water 2020 https://www.kaggle.com/vbmokin/phosphate-prediction-in-river-water

[11] Petruzevich D 2019 Time series forecasting using high order Arima functions *Proc XIX Int Multidisciplinary Scientific GeoConference SGEM-2019* 19(2.1) 673-80 doi: 10.5593/sgeom2019.2.1/S07.088

[12] Kalugin T R, Kim A K and Petruzevich D A 2020 Analysis of the high order ADL(p, q) models used to describe connections between time series *Russ Tech J* 8(2) 7-22 doi: 10.32362/2500-316X-2020-8-2-7-22

[13] Granger C W J 1980 Testing for causality: A personal viewpoint *J of Economic Dyn and Control* 2 329-52

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

[15] Hatemi J A 2008 Tests for cointegration with two unknown regime shifts with an application to financial market integration *Empirical Economics* 35(3) 497–505 doi:10.1007/s00181-007-0175-9

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

[17] Nkoro E and Uko A K 2016 Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation *J of Stat and Econometric Methods* 5(4) 63-91

[18] Pesaran M H and Shin Y 1999 *An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis* (Cambridge: Cambridge University Press, Cambridge. Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium)

[19] James G, Witten D, Hastie T and Tibshirani R 2015 *An introduction to statistical learning with applications in R* (Springer-Verlag New York) doi: 10.1007/978-1-4614-7138-7

[20] Hastie T, Tibshirani R and Friedman J 2009 *The elements of statistical learning* (Springer-Verlag New York)