Vol. 6, No. 1, 2024

Mykhailo Bordun[1], Olha Mokrytska[2]

[1] Information Systems Department, Ivan Franko National University of Lviv, Ukraine, Lviv, University Street, 1, E-mail: Mykhailo.Bordun@lnu.edu.ua, ORCID 0000-0001-8818-3363

[2] Computer Design Systems Department, Lviv Polytechnic National University, Ukraine, Lviv, S. Bandery street 12, E-mail: olha.v.mokrytska@lpnu.ua, ORCID 0000-0002-2887-9585

DEVELOPMENT OF SOFTWARE AND ALGORITHMIC EQUIPMENT FOR PREDICTION OF RIVER WATER POLLUTION USING FRACTAL ANALYSIS METHODS

Received: March 11, 2024 / Revised: April 01, 2024 / Accepted: April 05, 2024

https://doi.org/

Abstract. This paper explores the application of the ARFIMA fractal model for prediction of the dynamics of river water pollution based on BOD measure. The study begins by conducting a review of related works in the field of water quality analysis. At this stage also a suitable dataset is selected, that is used to train the ARFIMA, one of the machine learning models. GPH semi-parametric algorithm is applied for estimating the fractal differentiation parameter of the ARFIMA. The obtained results are compared with similar obtained with ARIMA model using RMSE and MAPE metrics. The study reveals an enhancement in accuracy with the use of fractal methods for water pollution prediction.

Keywords: fractal model, ARFIMA, biochemical oxygen demand, utoregressive model, ARIMA, Python, R language.

Introduction

Human economic activity causes significant changes in the environment. The water environment is not only a place where various pollutants accumulate, but due to the movement of water both over the land surface and through underground streams, these pollutants are spread across the planet. Changes in the physical, chemical and biological parameters of natural waters lead to negative consequences, the first of which is the harmful impact on human health and other living organisms. Water pollution is primarily caused by industrial, municipal and agricultural wastewater, the total volume of which is around 1,300 km^3 globally. At the same time, pollutants are released into the aquatic environment, the total weight of which is about 15 billion tonnes per year [1]. The key pollutants are heavy metals, dyes, pesticides, oil and oil products, biogenic organic matter, surfactants, etc. In addition to water-soluble pollutants, water is also contaminated with mechanical impurities, insoluble debris, and thermal and biological pollutants. It should be noted that water quality in rivers is deteriorating not only due to human activity but also due to natural factors. These include rock weathering, water evaporation, atmospheric deposition, climate change and natural disasters [2-4]. Both anthropogenic and natural impacts on water quality have certain seasonal changes and depend on the territory (urban or rural). Therefore, understanding what changes and factors affect river water quality is crucial for managing water quality in river basins, as well as predicting these changes in the future.

To achieve this goal, the following tasks were performed:

1. To select a historical data set for the selected river and analyse a limited sample for a specific study station to understand general trends and properly clean the data for further steps.
2. Decompose the time series and analyse it for white noise, stationarity and long memory.

3. Taking into account the results in the second step, select the most appropriate parameters of the selected fractal model to maximise the accuracy in terms of the RMSE metric.

4. Use the semiparametric GPH algorithm to estimate the fractal differentiation parameter of the ARFIMA model.

5. Compare the performance of the created fractal model with the same model with automatic parameter selection, as well as with the most appropriate autoregressive model on different sizes of training and test data. Analyse the results.

The object of the study is the dynamics of water pollution, which must be determined from previous historical data.

The subject of the study is the mathematical and software of an intelligent system for analysing and predicting the dynamics of water pollution.

The scientific innovation consists of the development of an intelligent system for the dynamics of analysis and forecasting, which will automate the data processing process and improve the quality of forecasting decisions, as well as the use of the semiparametric GPH algorithm to estimate the fractal differentiation parameter of the ARFIMA model.

The practical significance of the work lies in the development of software, and the functioning of an information system using fractal modelling approaches, which will facilitate the early detection of environmental problems.

Problem Statement

The subject area is rivers, which are complex and dynamic systems. The assessment of their quality is determined by many factors.

A common indicator of organic pollution of rivers is an indicator called biochemical oxygen demand (BOD) [2]. This is the amount of dissolved oxygen consumed by microorganisms during the oxidation of organic matter in water and waste. Typical sources of BOD are readily degradable organic carbon and ammonia. These compounds are common constituents or by-products of the metabolism of plant and animal waste and human activities (domestic and industrial wastewater). Standardised methods for quantifying BOD in wastewater have remained virtually unchanged for decades, despite numerous drawbacks. The most commonly used indicator is BOD_5, which determines the amount of oxygen in milligrams required for the oxidation of organic matter contained in 1 litre of water by aerobic bacteria to CO_2 and H_2O within 5 days without access to air and light. As a rule, the five-day period is not sufficient for complete oxidation, but it provides sufficient time for microbes to acclimatise and for significant (approximately 40 to 80 per cent) oxidation [2]. The cleanest rivers have a BOD_5 value of less than 1 mg O_2/l, while moderately and heavily polluted rivers have values between 2 and 8 mg O_2/l.

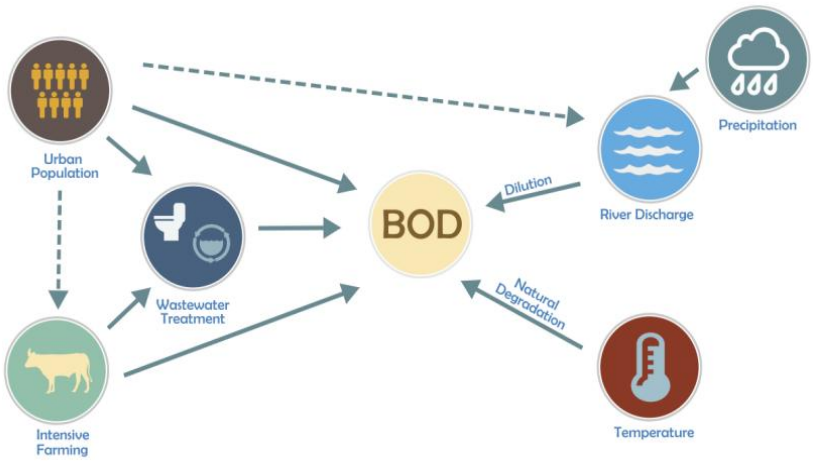

Fig. 1. Variables and processes affecting organic pollution in rivers, expressed as BOD [3]

The level of organic pollution in a river, usually expressed as BOD, is the result of two opposing mechanisms: pollutant inputs and natural purification (Figure 1) [3]. Wastewater discharges from cities and intensive livestock farms are the main organic pollutants in rivers. Although the pollution originates at the point of discharge along the river, the impact of such pollution extends to downstream populations and ecosystems as pollutants are transported through the river network. The extent to which pollutants are exposed downstream depends on the ability of rivers to clean themselves through dilution by natural runoff and natural degradation by microorganisms. Changes in river flow as a result of global warming affect the ability of rivers to dilute, especially in places where there is a decrease in climate humidity. An additional negative factor is the increase in water extraction to support the population. This further reduces the ability of rivers to function [3].

Despite the significant self-purification capacity of rivers, the number of people affected by organic pollution is projected to increase from 1.1 billion in 2000 to 2.5 billion in 2050 [3]. That requires not only the introduction of new water purification systems but also additional ways to predict changes in water quality indicators.

Review of Modern Information Sources on the Subject of the Paper

Time series analysis refers to the identification of general patterns reflected by data over a certain period of time. Among the most popular methods for modelling and forecasting time series are autoregressive models, the most prominent of which is ARIMA (AutoRegressive Integrated Moving Average). During the modelling process, we find 3 parameters [4]: Autoregressive (AR) of order p, which describes the number of significant delays in the time series, the order of integration of the series d, and the Moving Average (MA) of order p, which describes the number of significant forecast errors. A generalisation of this model is the fractal model ARFIMA (AutoRegressive Fractionally Integrated Moving Average), which allows modelling time series with a long memory. That is, the d-index can take on non-integer values. In general, this type of generalisation allows for the necessary analysis of various time series, taking into account a long-lasting shock in the time series, since the ARIMA model allows for the recording of processes with either a short memory with $d = 0$ or an infinite memory with $d = 1$. In his study [5] comparing the modelling of stationary time series using ARMA and ARFIMA processes, Anderson (1998), using Monte Carlo simulations and comparing forecast errors, shows that ignoring long memory when it exists leads to a more seriously deterioration of results than imposing it in the absence of it. This observation is extremely important because, in practice, a researcher never knows which process underlies the dynamics of processes. Based on the above considerations, we can consider the use of ARFIMA processes as one of the most modern and relevant approaches for studying time series.

Analysis of software tools.

The studies related to time series analysis and fractal differentiation were carried out with the Python programming language version 3.6.5 using the pandas version 1.1.3 and numpy version 1.19.2 libraries. For time series forecasting, the R programming language version 4.1.3 was used, along with the forecast version 8.16 and Arfima version 1.8.0 libraries. The division into two programming languages for the implementation of this work was forced since there are no libraries related to working with the ARFIMA model or other fractal models in such popular programming languages as Python.

Main Material Presentation

Dataset for forecasting river water pollution

The dataset has been taken from the data.gov.uk | Data publisher platform [6]. The data for its creation has been obtained between 1990 and 2018. This table presents raw data from river monitoring sites, including Water Framework Directive monitoring sites.

The dataset consists of a single CSV file with aggregated data for various rivers in the UK, each value representing a measurement from a specific monitoring station at a particular time for a particular river. The data for each measurement has been taken at different intervals of 1-2 times per month, and are

represented by several characteristics, such as alkalinity, biochemical oxygen demand, conductivity, dissolved copper, dissolved oxygen, dissolved iron, nitrate, nitrite, ammonia, acidity, dissolved phosphorus, suspended solids, and dissolved zinc.

The river with the highest number of measurements from this dataset, namely the River Quoyle (Northern Ireland), was selected for the study (458 measurements). The BOD was chosen as it is a key characteristic for assessing river pollution and is often used as a very important indicator for pollution assessment in the relevant scientific literature. However, preliminary data cleaning was required, as not every measurement had this indicator.

Investigating the presence of trend, seasonality, Hurst's indicators, stationarity of time series and long memory

Before starting modelling and forecasting, it is necessary to identify certain properties of time series to understand the feasibility of using specific machine learning models.

Thus, the first step is to decompose the time series, which can be seen in Figure 2, created using the statsmodels.api module of the Python programming language.

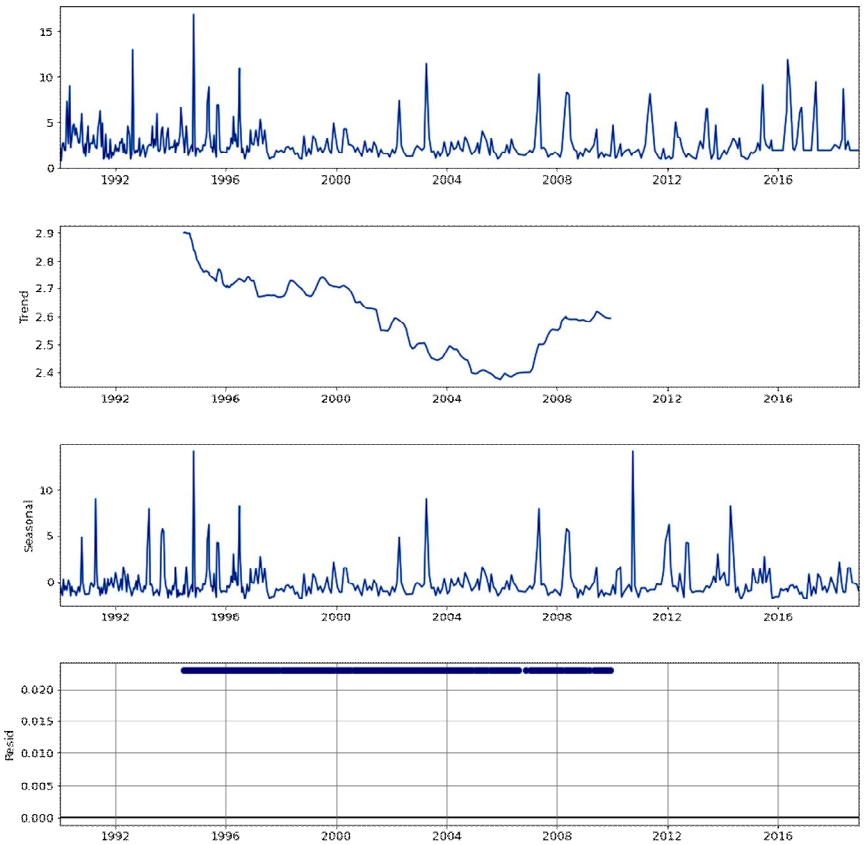

Fig. 2. Decomposition of the time series of BOD in the Quoyle River

It shows the presence of a trend and the absence of pronounced seasonality. The last graph in Figure 2 characterises the outliers, i.e. shows the noise in our time series. Ideally, it should have white noise patterns, which are characterised by zero mean, constant variance, and independence of variables. In other words, it is actually a random set of numbers, and this allows machine learning models to capture all the necessary signals of the training sample, which maximises the efficiency of the algorithm. However, in our case, we can see that although the variance is stable, the mean value is always greater than zero, and we can safely reject the hypothesis of white noise in the redundancy graph.

The next step in the analysis is to find the Hearst index, which is one of the most popular methods for assessing memory in a time series. The author of this method, Harold Hirst, was the first to explore the

concept of long memory when he studied the tributaries of the Nile River and the optimal size of water reservoirs. In general, the value of this indicator ranges from 0 to 1. The value of the Hearst exponent for the time series was determined using R/S analysis [7]. In this case, we obtained a value of 0.6266, which indicates a moderate long-term correlation in the time series with a tendency to randomness.

Figure 3 also shows the ACF graph, i.e. the autocorrelation graph, which shows the correlation between the time series and its lagged version at each value of the delay. The graph shows that some time delays fall outside the confidence interval (blue area on the graph), which expresses the dependence and point effect of long-term memory.

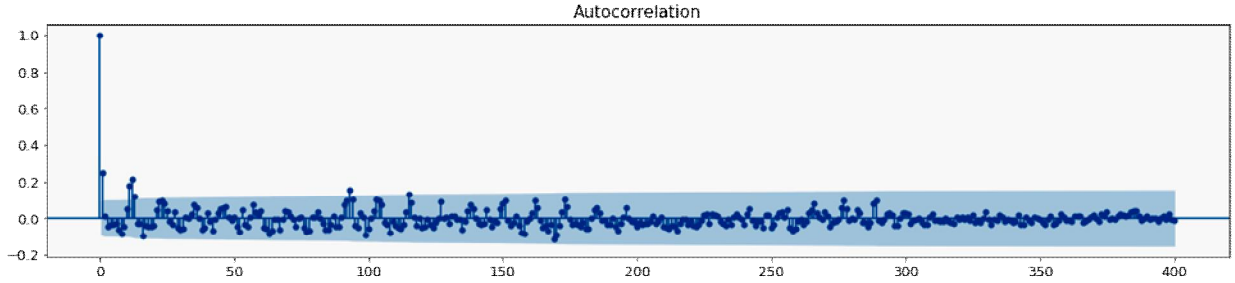

Fig. 3. ACF function for the time series of BOD in the Quoyle River

There are two popular statistical tests for determining stationarity: ADF and KPSS [8-10]. In general, each of them tests opposite hypotheses and results of which are complementary to each other, i.e. the impossibility of rejecting the null hypothesis of non-stationarity by the ADF test and rejecting this hypothesis of stationarity by the KPSS test is an important indicator of the presence of a unit root, i.e. an I(1) process, which shows non-stationarity of the time series.

The result was obtained using the statsmodels.tsa.stattools module of the Python programming language, which showed a p-value of the ADF statistic of 0.00014, which is less than the threshold of 0.05, which serves to reject the null hypothesis. The p-value of the KPSS statistic is 0.0751, which is greater than the threshold of 0.05 and accepts the null hypothesis for our time series. Stationarity is a rather important attribute and makes the time series more predictable and suitable for applying various forecasting methods.

Mathematical description of the fractal model ARFIMA

Autoregressive (AR) model of order p

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t \,, \tag{1}$$

where $\phi_1,...,\phi_p$ are autoregressive parameters, c is a constant and the random variable and ε_t is white noise.

Moving Average (MA) model of order q:

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \,, \tag{2}$$

where $\theta_1,...,\theta_p$ are parameters of the moving average, μ is constant and random variables ε_t, ε_{t-1}, ... is white noise.

A generalisation of the above models is:

$$(1 - \sum_{i=1}^{p} \phi_i B^i)(1-B)^d (X_t - \mu) = \left(1 + \sum_{i=1}^{q} \theta_i B^i\right)\varepsilon_t \,, \tag{3}$$

where $(1-B)^d$ is called the differentiation operator. The ARMA and ARIMA models can only capture processes with short memory since the parameter d takes on only integer values. Therefore, in order to capture processes with long memory, it is necessary to use the fractal model ARFIMA(p,d,q), where the parameter d takes on fractional values.

We can also expand the differentiation operator using the binomial expansion for any real number d:

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = 1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \ldots, \quad (4)$$

It is also possible to represent the differentiation operator as a Gamma function:

$$(1-B)^d = \sum_{k=0}^{\infty} \left(\frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}\right)(-B)^k, , \quad (5)$$

where Γ(.) denotes the Gamma function and is represented by $\Gamma(z) = \int_0^{\infty} x^{z-1}e^{-x}dx$. When d=0, X_t it is simply white noise, and its autocorrelation function is 0. When d=1, X_t is a random walk with the value of the autocorrelation function equal to 1, and it can be considered as white noise after first-order differentiation, which is the difference between the previous value and the current value in the time series. When d is already a real number, $X_t = -\sum_{k=1}^{\infty}\left(\frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}\right)X_{t-k} + \varepsilon_t$ and, therefore, X_t is influenced by all historical data (X_{t-1}, X_{t-2}, …).

Estimation of the optimal parameter of fractal differentiation of the ARFIMA model

Various methods for estimating the fractal differentiation parameter are increasingly mentioned in scientific literature related to time series analysis. These methods can be classified into two groups: parametric and semi-parametric. The most popular methods in the parametric group include a likelihood function. In the semi-parametric group, the most popular method, known as the GPH method, was proposed by Geweke and Porter-Hudak. The semiparametric approach involves estimating the model parameters in two stages. Firstly, the parameter d is estimated separately, and then the other parameters are estimated. This is different from the methods used in the first group, where all parameters are estimated simultaneously.

As the ARFIMA model from the R programming language forecast library with automatic parameter selection according to the documentation also uses a semi-parametric approach, we will provide a more detailed description of the GPH method.

The GPH method begins by estimating the parameter d through a least squares regression in the spectral domain, utilizing a sample shape from the spectral density poles at the origin.

$$f_X(\lambda) \sim \lambda^{-2d}, \quad \lambda \to 0.$$

To illustrate this method, we can express the spectral density function of the stationary model $X_t, t = 1,...,T$ as follows:

$$f_X(\lambda) = \left[4\sin^2\left(\frac{\lambda}{2}\right)\right]^d f_\varepsilon(\lambda), \quad (6)$$

where $f_\varepsilon(\lambda)$ is the spectral density of ε_t, assuming that it is a finite and continuous function on the interval $[-\pi;\pi]$.

The logarithm of the spectral density function can be represented as follows:

$$\log\{f_X(\lambda)\} = \log\{f_\varepsilon(0)\} - d\log\left\{4\sin^2\left(\frac{\lambda}{2}\right)\right\} + \log\left\{\frac{f_\varepsilon(\lambda)}{f_\varepsilon(0)}\right\}, \quad (7)$$

Let $I_X(\lambda_j)$ be the periodogram performed on the Fourier frequencies, $\lambda_j = \frac{2\pi_j}{T}, j = 1,2,...,m$.

T is the number of studies, and *m* is the number of Fourier frequencies considered, which is the number of ordinates of the periodogram to be used in the regression.

$$\log\{I_X(\lambda_j)\} = \log\{f_\varepsilon(0)\} - d\log\left\{4\sin^2\left(\frac{\lambda_j}{2}\right)\right\} + \log\left\{\frac{I_X(\lambda_j)}{f_X(\lambda_j)}\right\} + \log\left\{\frac{f_\varepsilon(\lambda_j)}{f_\varepsilon(0)}\right\}, \quad (8)$$

where $\log\{f_\varepsilon(0)\}$ is a constant, $\log\left\{4\sin^2\left(\frac{\lambda_j}{2}\right)\right\}$ is an exogenous variable and $\log\left\{\frac{f_\varepsilon(\lambda_j)}{f_\varepsilon(0)}\right\}$ is an uncertain.

The importance of the choice of m is evident as it significantly affects the estimation results. On the one hand, m should be small enough to consider only frequencies close to zero. On the other hand, m should be large enough to ensure convergence of the least squares estimate.

The GPH estimation requires two main assumptions related to the asymptotic behaviour of equation (8):

H1: for low frequencies, we assume that $\log\left\{\frac{f_\varepsilon(\lambda_j)}{f_\varepsilon(0)}\right\}$ is insignificant

H2: the random variables $\log\left\{\frac{I_X(\lambda_j)}{f_X(\lambda_j)}\right\}, j = 1,2,...,m$ the random variables are independent and identically distributed (IID) random variables.

Under hypotheses H1 and H2, we can write a linear regression

$$\log\{I_X(\lambda_j)\} = \alpha - d\log\left\{4\sin^2\left(\frac{\lambda_j}{2}\right)\right\} + e_j, \quad (9)$$

where $e_j \sim н.о.р.(-c, \frac{\pi^2}{6})$. Let $Y_j = -\log\left\{4\sin^2\left(\frac{\lambda_j}{2}\right)\right\}$. The GPH estimate is the least squares estimate of the regression $\log\{I_X(\lambda_j)\}$ on the constants and α and Y_j . The estimate of d, denoted as $\hat{d}_{GPH}$ is defined as follows:

$$\hat{d}_{GPH} = \frac{\sum_{j=1}^{m}(Y_j - \bar{Y})\log\{I_X(\lambda_j)\}}{\sum_{j=1}^{m}(Y_j - \bar{Y})^2} \quad (10)$$

where $\bar{Y} = m^{-1}\sum_{j=1}^{m}(Y_j)$ and $m = g(T)$ with $\lim_{T\to\infty} g(T) = \infty$, and $\lim_{T\to\infty} g(T)/T = 0$.

Geweke and Porter-Hudak have shown that if $T \to \infty$ and $|d| < \frac{1}{2}$ we have

$$\sqrt{m}\left(\hat{d}_{GPH} - d\right) \sim N\left[0, \frac{\pi^2}{6}\left\{\sum_{j=1}^{m}(Y_j - \bar{Y})^2\right\}^{-1}\right]. \quad ((11)$$

Porter-Hudak (1990), Crato and de Lima (1994), showed that the parameter m should be chosen so that $T \to \infty$, $m = T^{\upsilon}$, $\upsilon = 0.5,0.6,0.7$. Under the normality assumption for X_t, they proved that the resulting estimate is consistent and asymptotically normal. Therefore, the estimated standard deviation can be used for inference.

Results and Discussion

The modelling and prediction task requires a dataset with the date of measurement and the corresponding BOD value.

To determine the optimal fractal model for ARFIMA, we first used the GPH method described in Section 3.4, using the fdGPH function of the fracdiff library of the R programming language. It was used

to estimate the fractal differentiation parameter. In order to obtain a stationary time series using the estimated fractal differentiation parameter, the time series were differentiated using the diffseries function of the fracdiff library of the R programming language, which uses an approximate binomial expression of the long memory filter. The determined differentiated series and the values of the ACF and PACF functions are shown in Figure 4.

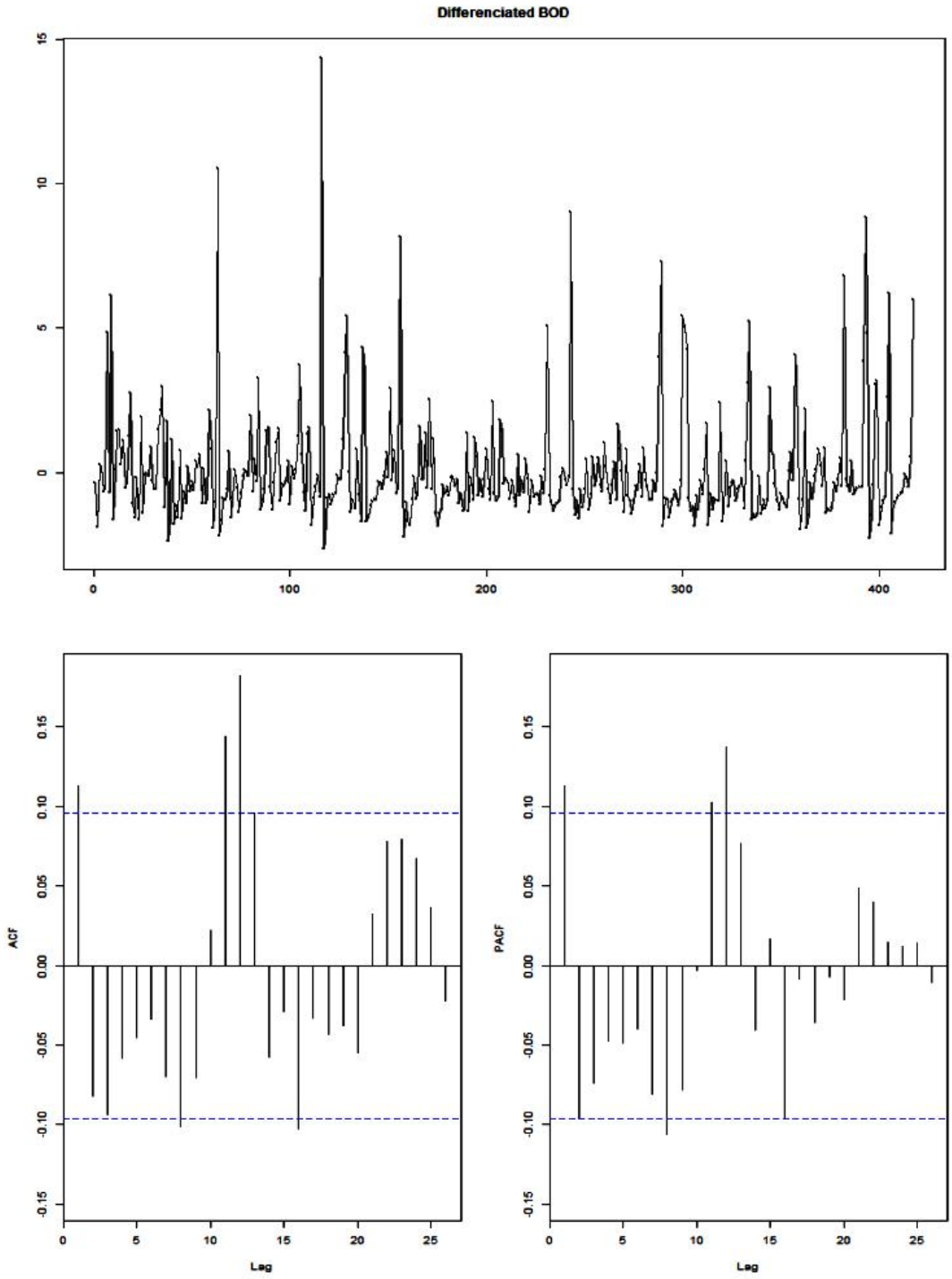

Fig. 4. Differentiated time series and applied ACF, PACF functions

From the values of the ACF and PACF functions, we can distinguish the number of significant delays, i.e. those that are greater than the range of the confidence interval (blue dashed line in Figure 4),

which will allow us to determine the maximum number for our parameters AR(p) from the ACF graph and MA(q) from the PACF graph, respectively.

After determining the possible values of the order of AR(p) and MA(q), we determined the d-index of the series integration using the fracdiff function from the fracdiff library of the R programming language. The choice of the optimal model was determined by the RMSE estimate, taking into account the difference between the test data.

After performing the above modelling, it is necessary to evaluate the predicted data, which can be visually represented, for example, as in Figure 5, and for a sufficiently small test sample, the result of such visualisation will be sufficiently demonstrative.

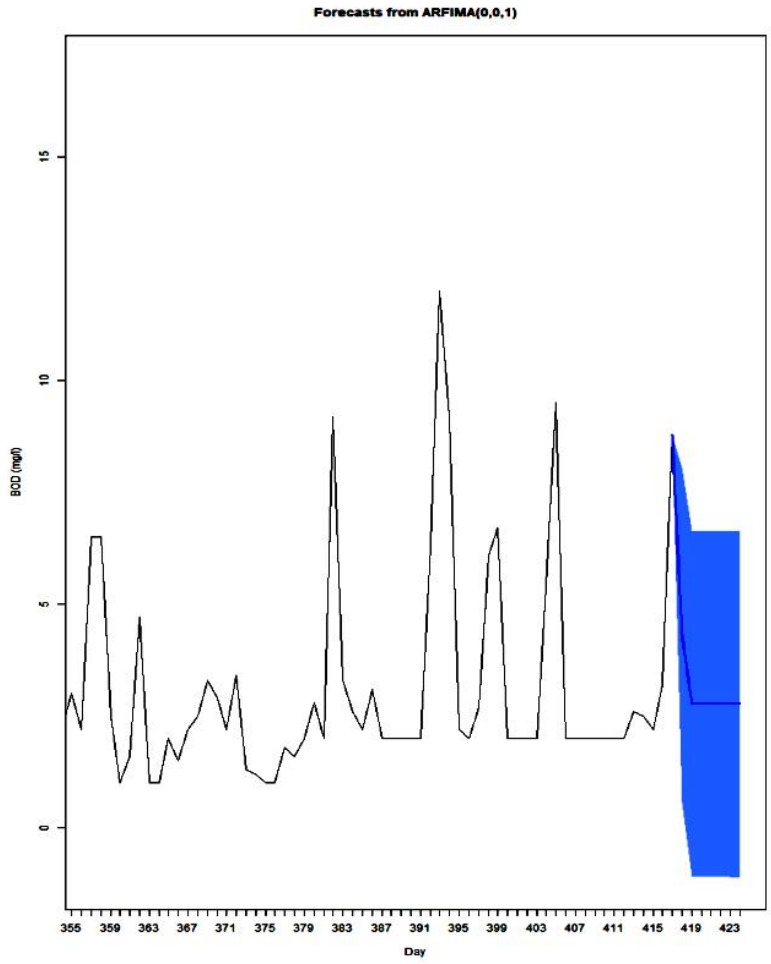

Fig. 5. Weekly forecast of BOD by ARFIMA(0,0,1) model

It is worth noting that, in addition to presenting the predicted data, the 95% confidence interval is also shown, which in Figure 5 is marked with a blue area and means the interval within which the value of the estimated random variable can be expected with a given confidence level.

Let's consider how well the trained network copes with the applied task of predicting the BOD index from the data for the River Quoyle (Northern Ireland). For this purpose, we use test data of 7 and 50 units, which were not used for training the network. In general, I compared the above models using two metrics for the test set, namely RMSE and MAPE. RMSE is the root mean square deviation between the predicted and actual values, while MAPE describes the average absolute percentage deviation. To better understand the effectiveness of using the ARFIMA fractal model, we also performed simulations using the ARFIMA and ARIMA models with automatic parameter selection from the R programming language's forecast library. The results are shown in Tables 1-2.

Table 1.

Evaluation of the prediction accuracy of test data between models (7 test data)

	RMSE	MAPE(%)
(auto) ARFIMA(0,0.004,1)	1.062	43.465
(auto) ARIMA(0,0,1)	1.059	43.363
manually fitted ARFIMA(0,0.165,0)	0.676	32.005

Table 2.

Evaluation of test data prediction accuracy between models (50 test data)

	RMSE	MAPE(%)
(auto) ARFIMA(0,0.003,1)	2.546	40.005
(auto) ARIMA(0,1,2)	2.607	33.620
manually fitted ARFIMA(13,0.1506,0)	2.526	36.351

According to the above results for the key metric RMSE, we can safely conclude that among the above models, the ARFIMA model with parameter estimation using the algorithm described in Section 5 has the highest forecasting efficiency. It is also important to note that despite the weakly expressed long memory in our time series, comparing the models with automatic parameter selection: fractal ARFIMA and autoregressive ARIMA, we can conclude that when there is a need to forecast a larger amount of data, the fractal model shows better results, since it takes into account the dependencies of previous values of the time series.

Conclusions

This paper discusses the use of information technology to forecast river pollution in non-stationary time series based on historical data.

The use of software tools to implement algorithms and forecasting models in Python is justified. The R programming language is used for forecasting time series. The study of Hurst's indicators, stationarity of time series and long memory is conducted. On the example of the River Quoyle (Northern Ireland), the results of forecasting on the full set of training data also confirm the feasibility of using the ARFIMA model in comparison with the ARIMA model, taking into account various estimates of the test error such as RMSE, MAPE, which is quite significant, since it demonstrates that even for stationary time series with minimal long memory, fractal models show better forecasting accuracy results. The algorithm for selecting the optimal parameter d of the fractal differentiation of the ARFIMA model is adapted.

In general, we can say that a larger number of data, both training and testing, clearly enhances the advantage of fractal models, as we take into account the long memory in the time series under consideration. However, you should always check the data and clean up any anomalies that cause an error in the forecasting estimate.

References

[1] S. K. Jain, V. P. Singh, Water resources systems planning and management. Elsevier, 2023

[2] M. R. Penn, J. J. Pauer, and J. R. Mihelcic, "Biochemical oxygen demand." Environmental and ecological chemistry, vol. 2, 2009, pp. 278-297. ISBN: 978-1-84826-206-5. – P. 278-297.

[3] Y. Wen, G. Schoups, and N. van de Giesen, Organic pollution of rivers: Combined threats of urbanization, livestock farming and global climate change. Sci Rep 7, 43289 (2017). https://doi.org/10.1038/srep43289.

[4] Liu K, Chen Y, Zhang X. An Evaluation of ARFIMA (Autoregressive Fractional Integral Moving Average) Programs. Axioms. 2017; vol. 6(2), P. 1-16. https://doi.org/10.3390/axioms6020016

[5] B. K. Ray, Modeling long-memory processes for optimal long-range prediction, Journal of Time Series Analysis, 1993, vol. 14: pp. 511-525. https://doi.org/10.1111/j.1467-9892.1993.tb00161.x

[6] River Water Quality Monitoring 1990 to 2018, 2022. URL: https://ckan.publishing.service.gov.uk/dataset/river-water-quality-monitoring-1990-to-201821

[7] D. Safitri, Mustafid, D. Ispriyanti and Sugito, Gold price modeling in Indonesia using ARFIMA method, IOP Conf. Series: Journal of Physics: Conference Series, vol. 1217, 2019, pp. 012-087, https://doi.org/10.1088/1742-6596/1217/1/012087

[8] V. Shah, G. Shroff Forecasting Market Prices using DL with Data Augmentation and Meta-learning: ARIMA still wins!, 2021, URL: https://arxiv.org/abs/2110.10233

[9] B. Mohamed, R. Khalfaoui, Estimation of the long memory parameter in non stationary models: A Simulation Study, 2011, URL: https://shs.hal.science/halshs-00595057

[10] V. Reisen, B. Abraham, S. Lopes, Estimation of parameters in ARFIMA Processes: A simulation study. 2006.

Михайло Бордун[1] , Ольга Мокрицька [2]

[1] Кафедра інформаційних систем, Національний університет ім. Івана Франка, Україна, Львів, вул. Університетська, 1, E-mail: Mykhailo.Bordun@lnu.edu.ua, ORCID 0000-0001-8818-3363

[2] Кафедра систем автоматизованого проектування, Національний університет «Львівська політехніка», Україна, Львів, вул. С. Бандери 12, E-mail: olha.v.mokrytska@lpnu.ua, ORCID 0000-0002-2887-9585

РОЗРОБЛЕННЯ ПРОГРАМНО-АЛГОРИТМІЧНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ ПРОГНОЗУВАННЯ ЗАБРУДНЕННЯ РІЧКОВИХ ВОД З ВИКОРИСТАННЯМ МЕТОДІВ ФРАКТАЛЬНОГО АНАЛІЗУ

Отримано: березень 11, 2024 / Переглянуто: квітень 01, 2024 / Прийнято: квітень 05, 2024

https://doi.org/

Анотація. У статті досліджено застосування фрактальної моделі ARFIMA для прогнозування динаміки забруднення річкових вод на основі вимірювання біохімічного споживання кисню . Дослідження починається з огляду суміжних робіт у галузі аналізу якості води. На цьому етапі також вибирається відповідний набір даних, який використовується для навчання ARFIMA, однієї з моделей машинного навчання. Напівпараметричний алгоритм GPH застосовано для оцінки параметра фрактального диференціювання ARFIMA. Отримані результати порівнюються з аналогічними, отриманими для моделі ARIMA з використанням метрик RMSE та MAPE. Дослідження виявило підвищення точності прогнозування забруднення води з використанням фрактальних методів.

Ключові слова: фрактальна модель, ARFIMA, біохімічне споживання кисню, авторегресійна модель, ARIMA, Python, мова R