TIME SERIES ANALYSIS OF WATER POLLUTION DATA

I. Drozdov

Russian Technological University (MIREA), Prospekt Vernadskogo, 78, 119454 Moscow, Russia

drozdov_i@mirea.ru

I. 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. Also, it has been shown that such models could be used for analysis and prediction of ecological data [1]. In particular, ARIMA models have been successfully applied to air pollution data around the world [2,3,4]. 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) [5], ammonium ions concentration [6] and phosphate ions concentration [7] 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 behavior 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.

II. DATA STRUCTURE

Three major indicators of water pollution are investigated: BOD (Biochemical Oxygen Demand), 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. There are values of seven water station measuring pollution level. Missing values in all analyzed time series have been handled by means of spline functions.

III. EXPERIMENTS

Autoregressive moving-average (ARIMA) models have been used to describe temporal dependencies in selected time series. Orders of ARIMA models have been chosen through exploratory analysis of ACF/PACF plots and numerical optimization of information criteria. Some of models showed quarter seasonality, which have been used to describe data using seasonal ARIMA (SARIMA) models.

Linear regression models of ammonium and phosphates ions concentration level in river water have been constructed. The models 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. 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 also shown.

IV. CONCLUSIONS

Exploratory data analysis detects irregular pollution spikes in the river water. ARIMA models have been used to describe how those spikes affect river water quality in the following periods and how river ecosystem responds to introduced external pollutants. Constructed ArDL models describe dependency between phosphates/ammonium and BOD, providing useful relations how pollutant ions affect BOD directly and over time. Coefficients of those model can be used to describe rates of those processes. Also, regression models fitted to spatially distributed data between water stations were able to predict where exactly pollutants have been introduced to the river and how long it affects quality of water further downstream.

This work is meant to show how simple and computationally inexpensive yet powerful and interpretable statistical methods could be used for analyzing environmental data to construct ecological mathematical models, which require only basic data that is currently collected due to governmental regulations. Alternative, more fundamental way of constructing such models would require extensive theoretical analysis of the problem, involving experts and field work; such models would generally be more correct and descriptive, but their construction is not always feasible.

REFERENCES

[1] Ives, Anthony & Abbott, Karen & Ziebarth, Nicolas. (2010). Analysis of ecological time series with ARMA (p, q) models. Ecology. 91. 858-71. 10.1890/09-0442.1.

[2] Abhilash, M. & Thakur, Amrita & Gupta, Deepa & Sreevidya, B.. (2018). Time Series Analysis of Air Pollution in Bengaluru Using ARIMA Model. 10.1007/978-981-10-7386-1_36.

[3] Rahman, Md Arafat & Hossain, M.A.. (2012). Time series analysis model for particulate matter of air pollution data in Dhaka city. Asian Journal of Water, Environment and Pollution. 9. 63-69.

[4] Guarnaccia, Claudio & Griselda, Cerón-Bretón & Breton, Rosa & Tepedino, Carmine & Quartieri, Joseph & Mastorakis, Nikos. (2018). ARIMA models application to air pollution data in Monterrey, Mexico. AIP Conference Proceedings. 1982. 020041. 10.1063/1.5045447.

[5] Prediction BOD in river water. Retrieved from (the 1st of October, 2020): https://www.kaggle.com/vbmokin/prediction-bod-in-river-water.

[6] Ammonium prediction in river water. Retrieved from (the 1st of October, 2020): https://www.kaggle.com/vbmokin/ammonium-prediction-in-river-water.

[7] Phosphate prediction in river water. Retrieved from (the 1st of October, 2020): https://www.kaggle.com/vbmokin/phosphate-prediction-in-river-water.

[8] Petrusevich D 2019 Time series forecasting using high order Arima functions *Proc XIX Int Multidisciplinary Scientific GeoConference SGEM 2019* **19(2.1)** pp 673-80 doi:10.5593/sgem2019/2.1/S07.088.

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

[10] I. Yu. Drozdov et al «Analysis of ammonium and phosphates concentration time series in river water», 2021 IOP Conf. Ser.: Earth Environ. Sci. 684 012021.

[11] I. Yu. Drozdov et al «Mathematical models of water pollution evaluation», 2021 IOP Conf. Ser.: Earth Environ. Sci. 684 012026.