ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 844 ACTA BIOLOGICA NR 21 2014

ANNA BUCIOR[*]
TYMOTEUSZ MILLER[*]
GORZYSŁAW POLESZCZUK[*]

APPLYING OF SEASONAL ARIMA MODEL FOR HYDROCHEMICAL MEASURING MISSING DATA COMPLEMENTATION

Abstract

An attempt to apply the seasonal ARIMA model to complement the missing one year hydrochemical data, particularly temperature, COD-Cr and total concentrations of phosphorus and iron on example of water outflow from the Rusałka Lake in city Szczecin (NW Poland).

Keywords: natural water, water quality, measuring results, ARIMA model, missing data complementation

Introduction

ARIMA models are investigative tools, which have been used to the analyses of experimental data in many fields of science (Demski 2004). With success ARIMA models are used to generating prognoses mainly in economic sciences (e.g. Snyder et al. 2001; Tseng et al. 2002; Andersen et al. 2003; Trzpiot and Orwat 2007; Talaga and Zieliński 1986; Talaga 1999). They found also application ingenerating

[*] Szczecin University, Faculty of Biology, Department of Chemistry and Natural Waters Ecosystems Management, Felczaka Str. 3c, 71-412 Szczecin, Poland, e-mail: polesz@univ.szczecin.pl.

prognoses e.g. in natural sciences (Montanari et al. 1997, 2000; Bucior et al. 2005; Poleszczuk et al. 2005) on the ground of long-term experimental data. Therefore the forecasting models are the most important to predicting data values.

In this work was apply attempt the ARIMA models to estimated missing experimental data (Talaga and Zieliński 1986; Talaga 1999) for one year in 7th-years investigated series. This test was realized for chosen surfaces water quality indices (water temperature, COD-Cr, $P_{tot.}$ conc. and $Fe_{tot.}$ conc.) on outflow waters of the Rusałka Lake in Szczecin (NW Poland) (Figure 1).

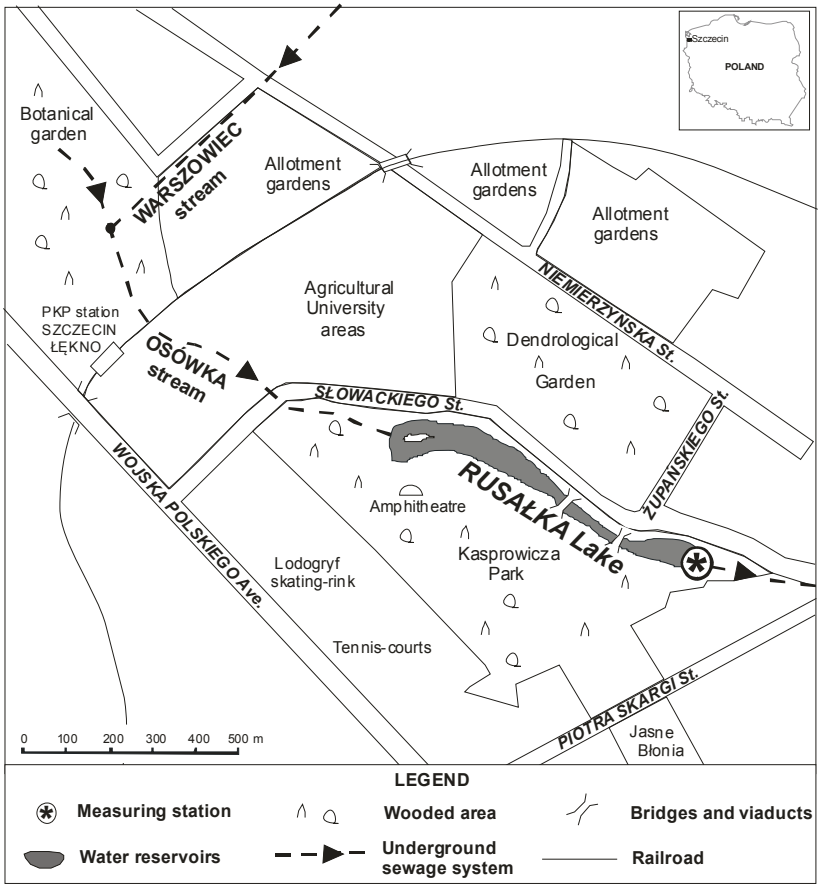

Figure 1. Rusałka Lake in Szczecin (NW Poland). Measuring station location in Kasprowicz Park (after Poleszczuk and Bucior 2009)

Characteristic of Rusałka Lake

The Rusałka Lake (Figure l), located in Szczecin city agglomeration, in the Kasprowicz Park is extended bed of Osówka stream. The volume of supplying water inflow, the Rusalka lake could have been seen rather as a stream water flood area but not a lake – which is shown by estimating the average water retention time in Rusałka on ca 30 days (Poleszczuk et al. 2012a). These lake possesses elongated shape with narrowing in his central part. Water inflow to the reservoir by underground channel and outflows by the same method. Rusałka Lake is connected to the municipal sewage system of Szczecin city, which fulfill role of retention water reservoir.

Material and methods

In Department of Chemistry and Natural Waters Ecosystems Management water quality indices were investigated on the water outflow from Rusałka Lake (Figure 1) in Szczecin city (NW Poland) in years 1999–2005. Selected results of these study (temperature, COD-Cr, $P_{tot.}$ and $Fe_{tot.}$) are presented in these paper. Water quality indices values were obtained in accordance with specific analytical procedures: COD-Cr – ISO 6060:1989, $P_{tot.}$ – PN-EN 1189:2000, chapter 6 and PN-EN ISO 6878:2006 P., $Fe_{tot.}$ – ISO 6332:1988. Some part of the research results were presented in works: Poleszczuk and Wawrzyniak (2002), Poleszczuk et al. (2012a, b), Bucior et al. (2013).

On basic of the dataset from years 1999–2002 and 2004–2005 was building the model which describer the changes of these water quality values in support about ARIMA model (Mongiało 1995), and made an data estimation (Lesińska et al. 1997) in year 2003. Evaluated data was verified by comparison with experimental data.

Results and discussion

All of the calculations were made in statistic applications contained within computer library Statistica 11 PL (Lesińska et al. 1997), time series and forecastings, ARIMA models. Choose ARIMA method to checking its usefulness to describing incomplete data (Talaga and Zieliński 1986; Talaga 1999) with regular and nonregular seasonal changes (Talaga and Zieliński 1986; Talaga 1999; Zawadzki and Goc 2010). The data

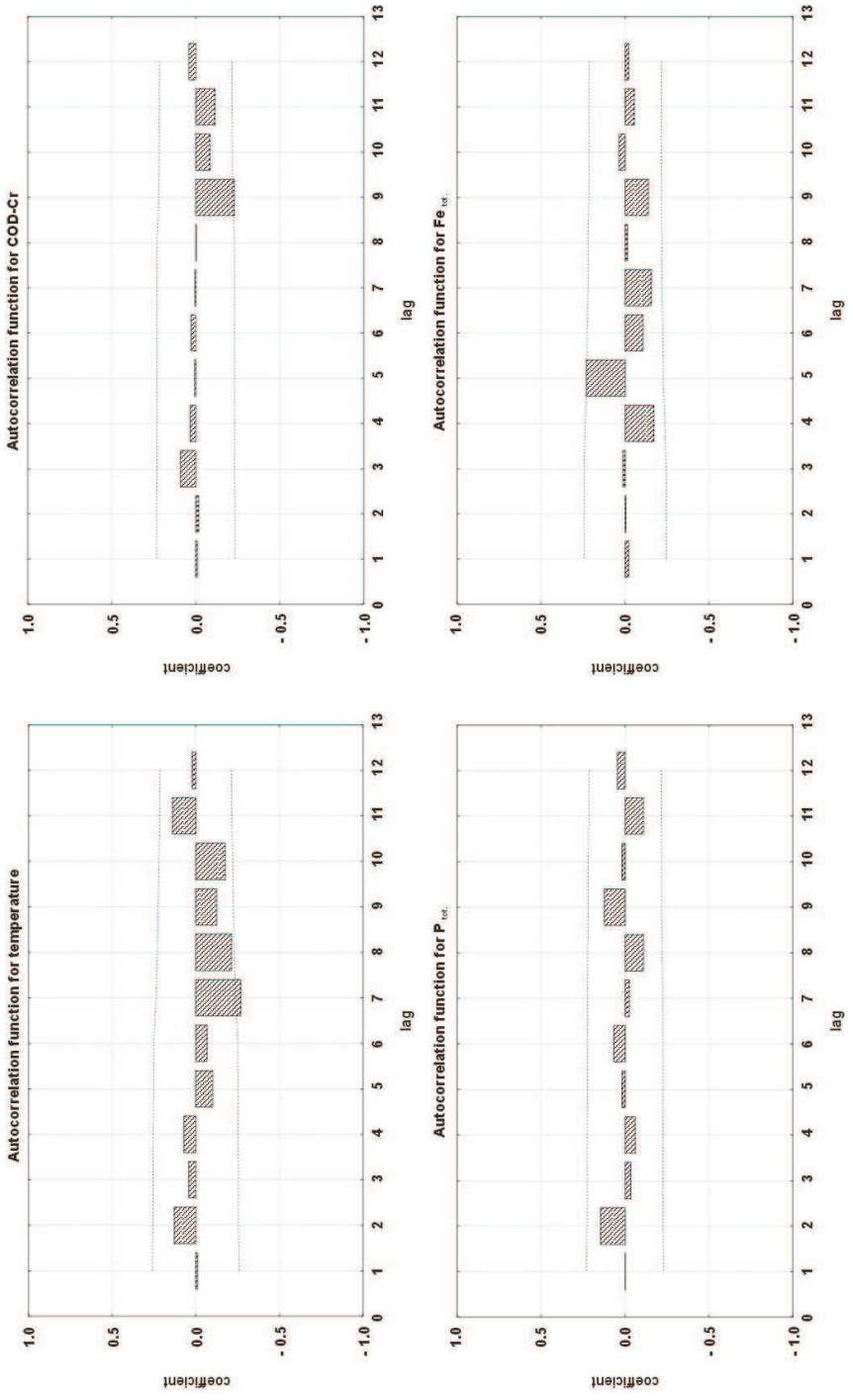

Figure 2. Autocorrelations on time series for selection waters quality coefficients

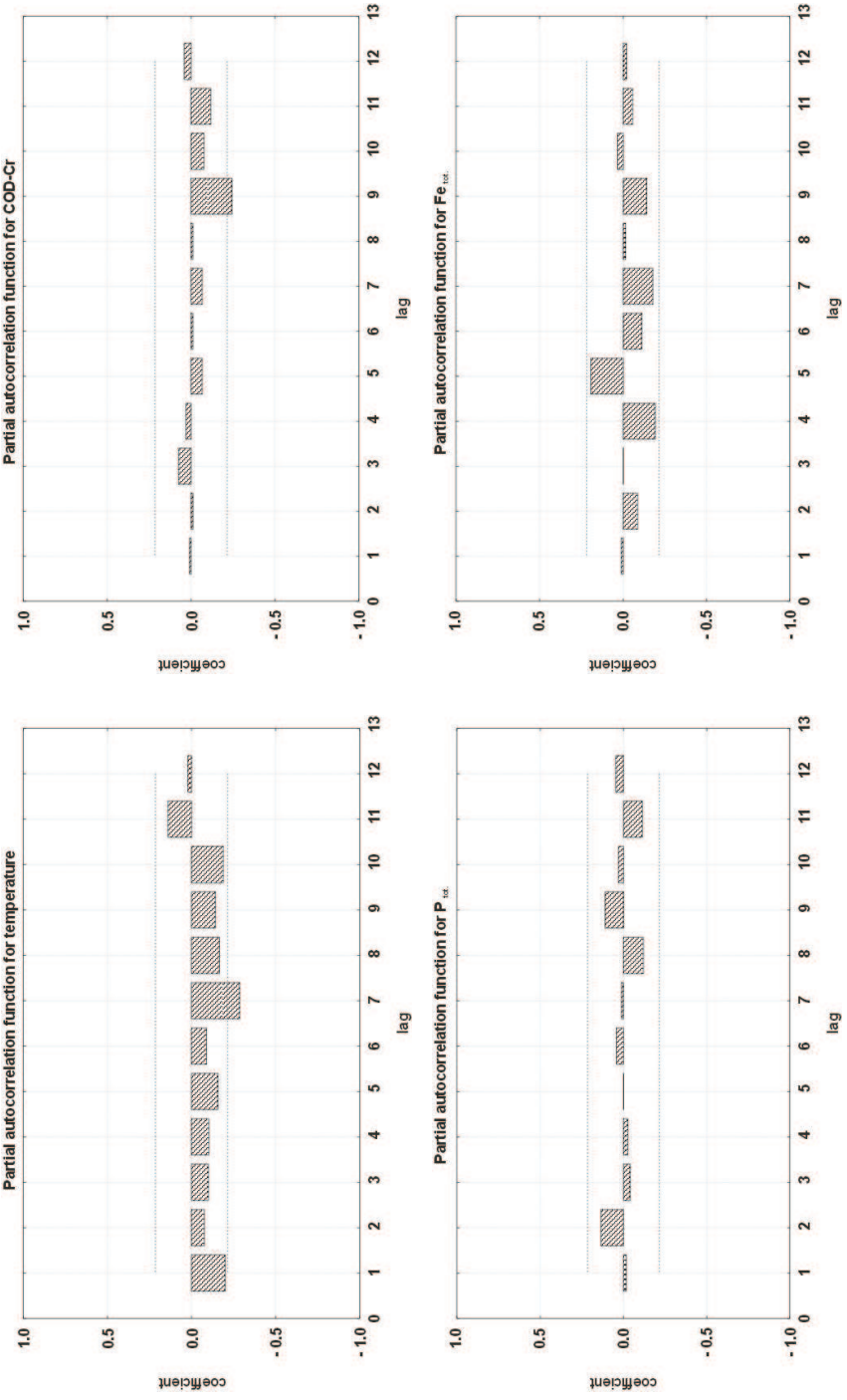

Figure 3. Partial autocorrelations on time series for selection waters quality coefficients

of recruitment samples and determinations chosen coefficients characterized water quality of surface waters (temperature, COD-Cr, $P_{tot.}$ and $Fe_{tot.}$) in years 1999–2005 was treated as time discreet series with one-year space. For so defined rows was marked autocorrelation (Figure 2) and partial autocorrelation (Figure 3) choosing ARIMA model with backwards intervention. To the calculations assume that the intervention is gradual and durable.

Repeated seasonable changes in year-old cycles were treated as autoregression model (p) of moving average (q) with time-lag one year (12 months) which describe time series, for which (p + q) > 3 (Andersen et al. 2003; Lesińska et al. 1997).

The Auto Regressive Integrated Moving Average (ARIMA(p, d, q)) models obtain to the forecasts incomplete data. For a given time series $\{x_n\}$, the persistence forecast is obtained by setting

$$x(n + 1) = x(n),$$

which implies that the average for water quality indices forecast for the next month.

The ARIMA models are traditionally very well suited to capture short range correlations, and hence have been used extensively in a variety of forecasting applications. Using the ARIMA models would require the inclusion of a large number of AR(p), MA(q) and differencing (d) parameters which would result in an expensive model.

Let $\{x_n\}$ represent the time series of monthly average for waters quality indices. Then, a ARIMA(p; d; q) formulation for the series can be described by equation:

$$\emptyset(B)\,(1 - B)^d (x_t - \mu) = \theta(B)\varepsilon_t,$$

where:

ε_t – is free term in the expression,

μ – is constant,

d – assumes fractional values.

B is the backshift operator defined by

$$B(x_t) = x_{t-1}.$$

The functions $\emptyset$, θ are polynomial functions of the backshift operator B:

$$\emptyset(B) = \varepsilon_t - \emptyset_1 B - \emptyset_2 B^2 - \ldots - \emptyset_p B^p \qquad (1)$$

$$\theta(B) = \mu + \theta_1 B + \theta_2 B^2 + \ldots + \varnothing_q B^q \tag{2}$$

where:

$\varnothing_1, \varnothing_2, \ldots$ – are parameters in autoregression (AR) model,

$\theta_1, \theta_2, \ldots$ – are parameters in moving average (MR) model.

The operator $(1 - B)^d$ is defined through a power series expansion:

$$(1 - B)^d = \sum_{j=o}^{j=\infty} \pi_j B^j ,$$

where $\pi_0 = 1$ and for $j > 0$

$$\pi_j = \prod_{k=1}^{k=j} (k - 1 - d)/k,$$

for $j = 1, 2, \ldots$

Thus, the ARIMA model is completely described by p parameters $\varnothing_p$ (where $p = 1, 2, ...$), q parameters $\varnothing_q$ (where $q = 1, 2, \ldots$) and the fractional parameter in equations (1) and (2). After defined the structure of the ARIMA model, the estimation of parameters in this model can be performed.

Then was executed estimation of parameters of this model McLeod's and Holanda-Sales method. Approximate McLeod's and Holanda-Sales is the fastest estimation method, and should be used in particular for very long time series, because not impose fixed limitations on the lengths of time series. McLeod and Holanda-Sales (McLeod and Holanda-Sales 1983) recommend to specify the number of backcasts so that:

$$\text{No. of backcasts} = q + s \cdot q_s + 20 \cdot (p + s \cdot p_s),$$

where:

p, p_s, q, and q_s – are the non-seasonal and seasonal autoregressive and moving average parameters, respectively,

s – is the seasonal lag.

Several iterations may be required to obtain convergence when the model contains moving average factor. Sufficient accuracy is usually obtained on the first step of evaluation.

Missing data to calculation were replacement by mean values from earlier data with interval 12 variables:

$$\bar{x} = \frac{\sum \frac{1}{\sigma_i^2} \cdot x_i}{\sum \frac{1}{\sigma_i^2}},$$

where:

$\bar{x}$ – mean values,
x_i – data value,
σ – interval (seasonal lag).

The next step was the test of construction of foreseeing model was undertaken missing in model values (Lesińska et al. 1997). The level of significance for studied model was accepted carrying out 95%. The corresponding parameters estimated for each water quality indices models was shown in Tables 1–4.

Table 1. Parameters of ARIMA model for water temperature

Parameter	$q_{(1)}$	$P_{s(1)}$	$Q_{s(1)}$	$Q_{s(2)}$	$Q_{s(3)}$
Estimation	0.00225	0.99999	0.58430	0.02778	0.11262

Table 2. Parameters of ARIMA model for COD-Cr

Parameter	$q_{(1)}$	$P_{s(1)}$	$Q_{s(1)}$	$Q_{s(2)}$	$Q_{s(3)}$
Estimation	–0.03400	0.99992	0.45734	–0.08190	0.23121

Table 3. Parameters of ARIMA model for conc. of $P_{tot.}$

Parameter	$q_{(1)}$	$P_{s(1)}$	$Q_{s(1)}$	$Q_{s(2)}$	$Q_{s(3)}$
Estimation	–0.04750	0.99994	0.63580	0.14146	–0.01120

Table 4. Parameters of ARIMA model for total Fe concentration

Parameter	$q_{(1)}$	$P_{s(1)}$	$Q_{s(1)}$	$Q_{s(2)}$	$Q_{s(3)}$
Estimation	–0.24700	1.0000	0.74528	0.16659	–0.13080

The changes of chosen quality coefficients in analysed period and its estimated values in missing 2003 year were presented on Figure 4 and on the Tables

5 and 6. The aim of verifying estimated values investigated coefficients, the value generated through chosen model ARIMA were compared from experimental data (Figure 4). From introduced on Figure 4 data, arise that investigated water quality indices showed reasonable, occurrent seasonal periodically changes (Bucior and Poleszczuk 2006; Talaga and Zieliński 1986; Talaga 1999; Zawadzki and Goc 2010).

Table 5. Forecasting values and experimental data for water temperature and COD-Cr

No.	Date	Water temperature		COD-Cr	
		forecasting (prognoses) values	experimental (observed) values	forecasting (prognoses) values	experimental (observed) values
48	January 2003	1.3	0.5	62	61
49	February 2003	9.0	7.0	68	69
50	March 2003	18.8	10.0	74	74
51	April 2003	14.7	14.0	65	73
52	May 2003	18.3	18.0	66	74
53	June 2003	20.5	22.0	61	70
54	July 2003	21.7	21.0	68	82
55	August 2003	21.3	20.0	73	73
56	September 2003	9.2	13.0	70	71
57	October 2003	6.9	8.0	84	86
58	November 2003	4.6	5.0	75	71
59	December 2003	1.9	0.0	72	76

Table 6. Forecasting values and experimental data for concentrations $P_{tot.}$ and $Fe_{tot.}$

No.	Date	$P_{tot.}$ conc.		$Fe_{tot.}$ conc.	
		forecasting (prognoses) values	experimental (observed) values	forecasting (prognoses) values	experimental (observed) values
48	January 2003	0.52	0.72	0.18	0.17
49	February 2003	0.53	0.45	0.22	0.25
50	March 2003	0.51	0.60	0.24	0.23
51	April 2003	0.51	0.50	0.18	0.55
52	May 2003	0.34	0.32	0.22	0.48
53	June 2003	0.91	1.19	0.18	0.53
54	July 2003	0.70	0.75	0.17	0.49
55	August 2003	0.85	1.39	0.19	0.38
56	September 2003	1.66	1.58	0.15	0.35
57	October 2003	1.00	1.22	0.23	0.45
58	November 2003	0.62	0.76	0.23	0.23
59	December 2003	0.48	0.50	0.18	0.15

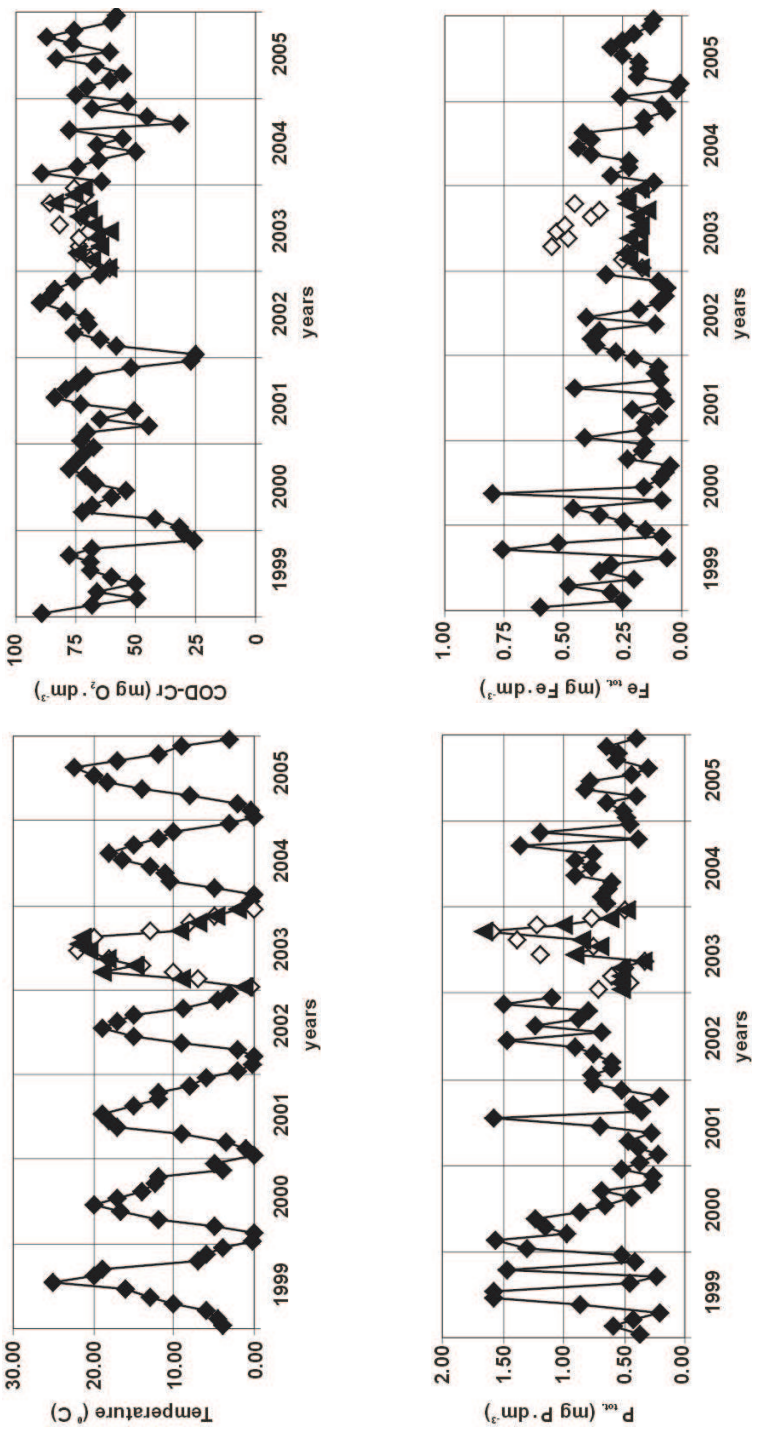

Figure 4. Observed values (—◆—) and their forecasts (—▲—) for selection waters quality coefficients. Experimental values (◇) was showed to compares with forecasting values

The water temperature values show very regular seasonal changes, therefore missing values had very good estimated through model ARIMA. Values of concentrations COD-Cr and $P_{tot.}$ show also seasonal changeability. However it was periodically irregular changeability. For these rows generated through model ARIMA data were satisfactory. The time series describe the changes of concentrations $Fe_{tot.}$ in surface waters outflow from Rusałka Lake show very irregular seasonal changeability. For these time series missing data generated through model ARIMA were compared from experimental data. It was affirmed, that for total Fe concentrations estimated values was unsatisfactory.

The models were applied to forecasts monthly chosen water quality indices in long-time series. The proposed method was not improve the accuracy of forecasting by the significance level of 95%.

Conclusions

ARIMA models can be used with success for estimating missing data (e.g. changes of water temperature in this work) in long-time series with very regular seasonal variability. In some special case (e.g. changes total concentration of Fe described in this work) when the seasonal changes are less regular – the ARIMA method is not the method which giving satisfaction results.

References

Andersen T.G., Boollreslev T., Diebold F.X., Labys P. 2003. Modelling and forecasting realized volatility. *Econometrica*, 71: 529–626.

Bucior A., Poleszczuk G., Wawrzyniak W. 2005. On possibility of the great Lagoon (Szczecin Lagoon, NW Poland) water quality prediction with the Arima Modeling support. Proc. VI Internat. Conf.: Analysis, forecasts and steering in matches systems, Ed. Tech. Univ. Sankt-Peterburg, Sankt-Peterburg, pp. 7–10.

Bucior A., Poleszczuk G. 2006. On possibility of predicting values of water quality indices in an estuary without tide – the Szczecin Lagoon (NW Poland) water quality prediction with the Arima Modeling support. Trudy VII Mieżdonarodnoj Nauczno – Prakticzeskaja Konf. Młodych Uczonych, Studentow i Aspirantow: Analiz i prognozirowanije sistem uprawlenija, 26–28 april 2006, Sankt-Peterburg, Izd. Siewiero-Zapadnyj Techniczeskij Uniwersytet (SZTR), Sankt-Peterburg, Part 1, pp. 25–31.

Bucior A., Miller T., Meller E., Wawrzyniak W., Poleszczuk G. 2013. Apply ARIMA models to forecasting long-term time series in natural sciences on the basic Rusałka

Lake in Szczecin city (NW Poland). Proceedings of 14th International Conference of Students, Postgraduates and young Professionals "Analysis and forecasting management systems", 23–26 April 2013 Sankt-Petersburg, Ed. State University of Communications in Saint Petersburg, Sankt-Petersburg, pp. 11–22.

Demski T. 2004. *Data mining in industry: projection, refining, production*. Ed. StatSoft Polska (in Polish).

ISO 6060:1989. Water quality – Determination of the chemical oxygen demand.

ISO 6332:1988. Water quality – Determination of iron – Spectrometric method using 1,10-phenanthroline.

Lesińska E., Sokołowski A., Wątroba J., Demski T., Jakubowski J. 1997. *Statistica PL for Windows*, Vol. III, Statystyki II, Ed. StatSoft Polska (in Polish).

McLeod A.I., Holanda-Sales P.R. 1983. Algorithm AS 191: An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. *Applied Statistics*, 32 (2): 211–232.

Mongiało Z. 1995. Predicting supply and demand of agricultural products using time series. In: *The integration food economy in western and northern Poland from European communities*. I. Rutkowska (ed.). Vol. II, Akademia Rolnicza in Szczecin, pp. 57–65 (in Polish).

Montanari A., Rosso R., Taqqu M.S. 1997. Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. *Water Resour. Res*., 33 (5): 1035–1044.

Montanari A., Rosso R., Taqqu M.S. 2000. A seasonal fractional ARIMA Model applied to the Nile River monthly flows at Aswan. *Water Resour. Res*., 36 (5): 1249–1259.

PN-EN 1189:2000, chapter 6. Water quality. Determination of total phosphorus using spectrometric method with ammonium molybdate after oxidation peroxydisulphate (VI) in the range of 0.01 mg/l (in Polish).

PN-EN ISO 6878:2006 P. Water quality – Determination of phosphorus – Spectrometric method using ammonium molybdate (in Polish).

Poleszczuk G., Bucior A. 2009. Biodegradation of organic substances in the waters of Lake Rusałka in Szczecin agglomeration. *Bulletin VURH Vodnany*, 45 (4): 44–51.

Poleszczuk G., Wawrzyniak W. 2002. Rusałka Lake in Szczecin – water quality investigations in four years after dredge bottom sediments. In: *Ekologia Pogranicza*. T. Zaborowski (ed.). Instytut Badań i Ekspertyz Naukowych, Gorzów Wielkopolski, pp. 163–167 (in Polish).

Poleszczuk G., Bucior A, Wawrzyniak W., Czerniejewski P. 2005. On possibility forecasting of the Roztoka Odrzańska waters quality (Odra river estuary, NW Poland) with the ARIMA Modeling support. In: *Ekologia Pogranicza-EP'05*. T. Zaborowski (ed.). Instytut Badań i Ekspertyz Naukowych, Gorzów Wielkopolski, pp. 432–439 (in Polish).

Poleszczuk G., Bucior A., Miller T., Tokarz M. 2012a. Pollution of the ecosystem of the Rusałka city lake with heavy metals. *Chem. Didact. Ecol. Metrol*., 17 (1–2): 75–88.

Poleszczuk G., Bucior A., Tokarz M., Pierwieniecki J. 2012b. Trophic status of the Rusałka Lake in Szczecin in years 1999–2010. *Acta Biologica*, 19: 37–48.

Snyder R.D., Ord J.K., Koehler A.B. 2001. Prediction intervals for ARIMA models. *Journal of Business and Economic Statistics*, 19: 217–225.

Talaga L. 1999. Models of stochastic processes and data gaps. In: *Econometric prediction methods for seasonal data in terms of the lack of complete informations*. J. Zawadzki (ed.). Rozprawy i Studia Uniwersytetu Szczecińskiego, 342: 141–176 (in Polish).

Talaga L., Zieliński Z. 1986. *Spectral analysis in econometric modeling*. PWN (in Polish).

Trzpiot G., Orwat A. 2007. Dynamic modeling of value units account OFF using ARIMA models. *Acta Universitatis Lodziensis*, 206: 279–298 (in Polish).

Tseng F.M., Yu H.C., Tzeng G.H. 2002. Combining neural network model with seasonal time series ARIMA model. *Technological Forecasting and Social Change*, 69: 71–87.

Zawadzki J., Goc M. 2010. Forecasting distributions of Variables with seasonal Fluctuations. Research Papers of Wrocław University of Economics, series *Econometrics*, 28: 195–207.

ZASTOSOWANIE SEZONOWYCH SZEREGÓW CZASOWYCH ARIMA DO UZUPEŁNIANIA NIEPEŁNYCH HYDROCHEMICZNYCH DANYCH POMIAROWYCH

Streszczenie

W artykule przedstawiono próbę zastosowania sezonowych szeregów czasowych ARIMA do uzupełniania braku jednorocznych pomiarowych hydrochemicznych danych, a w szczególności temperatury, ChZT-Cr oraz ogólnych stężeń fosforu i żelaza na przykładzie wód odpływających z jeziora Rusałka w Szczecinie (NW Polska).

Sowa kluczowe: wody naturalne, jakość wody, wyniki badań, modele ARIMA, uzupełnianie brakujących danych

Cite this article as: Bucior A., Miller T., Poleszczuk G. 2014. Applying of seasonal ARIMA model for hydrochemical measuring missing data complementation. *Acta Biologica*, 21: 23–35.