Environment Protection Engineering

Vol. 10 1984 No. 4

RYSZARD KRASNODĘBSKI*

IS THE REGRESSION METHOD A GOOD TOOL FOR WATER QUALITY PREDICTION?

The aim of the paper is to examine the regression forecasting verifiability of pollutant's (dissolved oxygen) load and, consequently, of its concentrations flowing through a river cross-section. Three types of models are verified, namely: linear, log-log and log-linear ones. The verification involves the following data taken from the Warsaw cross-section of the Vistula river: (i) ammonium nitrogen, iron and manganese, 14 years, (ii) dissolved oxygen, 10 years. The hypotheses about the consistency of the empirical distributions of loads and flow intensities with the suitable theoretical ones (the Kolmogorov test) have been also verified. The parameters of the regression lines have been computed for each year and the loads defined by the regression lines for a selected flow intensity. From the verification results listed in 3 tables it follows that from year to year when there occur the same flow intensities, the parameters of regression lines and loads flowing through the river cross-section show unsatisfactory instability.

1. The aim of this paper is to examine the regression forecasting verifiability of pollutant (dissolved oxygen) load values flowing through a river cross-section. We will verify three types of models:

$$\begin{aligned} &\text{model I: } L = a+bQ+\varepsilon,\\ &\text{model II: } \log L = a+b\log Q+\varepsilon, \qquad (1)\\ &\text{model III: } \log L = a+bQ+\varepsilon \end{aligned}$$

where Q denotes flow intensity, L – pollutant (dissolved oxygen) load, a and b are regression line coefficients, ε is the residual variable.

The verification of the models is based on the data taken from the cross-section Warsaw of the Vistula river. The data storage include the concentrations

* Institute of Mathematics, Technical University of Wrocław, pl. Grunwaldzki 13a, 50-378 Wrocław, Poland.

of: (i) ammonium nitrogen, iron and manganese observed during the years 1965–1978, (ii) dissolved oxygen observed during the years 1967, 1968, 1970, 1971, 1973–1978. The number of the data for each year vary (several exceptions see tab. 3) from 264 to 365.

Practical needs force sometimes the application of a mathematical method, even when the assumptions justifying the adduction of the theorems of the appropriate theory are not sufficiently accurately fulfilled. The tables presented in this paper show, however, that the divergence between the assumptions and their realizations results in so strong a prediction unverifiability that the significance of the models becomes rather unreliable.

2. Every regression line, as it is determined in the branch of our interest, represents a set of flow intensities and pollutant (dissolved oxygen) loads observed in a year, say R. One supposes, however, in practice, that it represents the years $R+1$, $R+2, \ldots$, too. Yet, there are no reasons to assume *a priori* that the set of the data taken at random in the year R will make a statistical sample representative of both the years R and $R+1$ or R and $R+2$, etc. In fact, combination of natural and economical phenomena produces relations between Q and L differing from year to year in such a way that the assumption mentationed above is quite frequently not justified (see tab. 2, columns 2, 5 and 8 and tab. 3).

3. We verify (tab. 1) the hypotheses which state that the probability distribution functions (pdf) of suitable random variables are normal or log-normal, respectively. We use Kolmogorov's test with the significance level 0.05; the critical value of the statistics is then equal to 1.358. The mean value of the annual sample and its variance have been assumed as parameters of the theoretical pdf.

As it is seen, neither normal pdf nor log-normal one are good approximations of empirical distributions for both flow intensities and loads. Table 2 shows almost the same inconsistency for residual variables. In practical use of the models I, II, and III, according to the knowledge of the author, these discrepancies are rather often neglected.

4. Table 3 contains a fairly significant verification. For each yesr, as far as it was possible, the following comparison has been made. We consider two loads:

(i) the load L_R "realized" in the year R when the flow intensity was equal to $1.2Q_R$ where Q_R is the mean value of loads observed in the year R; by the load "realized" in the year R we mean the load calculated from the regression function valid for the year R; consequently, L_R is considered as a prognosis load for the year $R+1$ for the moment when the flow intensity is equal to $1.2Q_R$,

(ii) the load L_{R+1} "realized" in the year $R+1$ when the flow intensity was equal to $1.2Q_R$; L_{R+1} is calculated from the regression function valid for the year $R+1$.

Table 1

Statistical parameters of flow intensities and loads of ammonia nitrogen, iron, manganese, and dissolved oxygen in the Vistula–Warsaw cross-section in the years 1965–1978

Parametry statystyczne natężeń przepływu i ładunków azotu amonowego, żelaza, manganu i rozpuszczonego tlenu w przekroju Wisła–Warszawa w latach 1965–1978

Year	Data number	Flow intensity, m³/s						Load, g/s					
		Mean value	Dispersion	The Kolmogorov test				Mean value	Dispersion	The Kolmogorov test			
				k_n		k_{ln}				k_n		k_{ln}	
1	2	3	4	5		6		7	8	9		10	
						Ammonia nitrogen							
1965	269	765	518	3.00	—	1.11	+[(2)]	337	936	5.89	—	5.72	—
66	331	844	504	2.77	—	1.21	+	152	215	4.36	—	7.26	—
67	294	691	511	3.24	—	2.75	—[(3)]	307	391	3.88	—	1.90	—
68	298	676	260	2.04	—	0.79	+	406	471	3.86	—	2.17	—
69	292	440	239	2.44	—	1.38	—	440	503	3.34	—	2.09	—
70	302	728	515	3.80	—	1.10	+	375	411	3.51	—	2.02	—
71	306	530	302	2.56	—	2.19	—	353	353	2.95	—	2.34	—
72	360	509	268	3.23	—	1.44	—	308	317	3.41	—	2.34	—
73	365	471	286	2.87	—	2.45	—	401	424	3.50	—	2.75	—
74	365	825	546	2.68	—	1.55	—	409	596	4.89	—	1.90	—
75	365	693	329	3.16	—	2.03	—	348	319	3.44	—	2.88	—
76	365	579	335	3.43	—	1.61	—	347	343	3.11	—	2.84	—
77	365	620	370	4.16	—	3.50	—	390	432	4.02	—	1.84	—
78	272	662	290	2.31	—	1.05	+	334	445	3.88	—	2.42	—
						Iron							
1965	272	761	517	3.02	—	1.10	+	3282	11911	6.46	—	5.30	—
66	331	844	504	2.77	—	1.22	+	3181	7796	6.22	—	6.13	—
67	293	693	512	3.22	—	2.73	—	2020	4083	5.44	—	2.33	—
68	298	676	260	2.03	—	0.79	+	1649	2639	4.79	—	0.95	+
69	292	440	239	2.45	—	1.38	—	1042	2125	5.51	—	3.17	—
70	303	727	514	3.40	—	1.09	+	2726	8750	6.59	—	1.23	+
71	306	530	303	2.56	—	2.19	—	1032	2059	5.60	—	1.89	—
72	361	511	269	3.17	—	1.40	—	966	3204	7.46	—	1.82	—
73	365	471	286	2.87	—	2.45	—	912	2509	6.92	—	2.23	—
74	365	825	546	2.68	—	1.55	—	1845	3794	6.07	—	1.07	+
75	365	693	329	3.16	—	2.03	—	1013	1616	5.48	—	2.48	—
76	365	579	335	3.43	—	1.61	—	1081	2308	6.27	—	1.80	—
77	365	620	370	4.16	—	3.50	—	941	1560	5.74	—	2.96	—
78	272	662	290	2.31	—	1.05	+	1313	2164	4.77	—	1.91	—

1	2	3	4	5		6		7	8	9		10	
						Manganese							
1965	264	773	520	2.99	–	1.14	+	192	416	5.23	–	5.53	–
66	331	844	504	2.77	–	2.23	–	193	397	5.70	–	6.60	–
67	291	693	513	3.20	–	2.71	–	162	214	4.98	–	2.33	–
68	298	676	260	2.03	–	0.79	+	185	219	5.03	–	2.04	–
69	191	431	252	2.05	–	1.01	+	177	172	2.50	–	0.56	+
70	303	727	514	3.40	–	1.10	+	237	607	6.06	–	2.96	–
71	306	530	302	2.56	–	2.19	–	121	179	5.64	–	2.20	–
72	361	510	269	3.20	–	1.42	–	117	212	6.16	–	0.95	+
73	364	470	286	2.87	–	2.45	–	105	128	4.55	–	1.69	–
74	364	826	546	2.67	–	1.55	–	146	243	5.69	–	2.12	–
75	365	693	329	3.16	–	2.03	–	92	82	3.97	–	4.40	–
76	365	579	335	3.43	–	1.61	–	142	163	5.41	–	2.42	–
77	365	620	370	4.16	–	3.50	–	176	178	5.03	–	2.23	–
78	272	662	290	2.31	–	1.05	+	238	325	4.51	–	1.46	–
						Dissolved oxygen							
1967	294	693	516	3.24	–	2.74	–	7467	5819	3.20	–	2.35	–
68	79	660	196	1.09	+	0.59	+	6064	1823	1.42	–	1.04	+
69													
70	89	767	529	2.16	–	1.12	+	8078	4986	1.23	+	1.15	+
71	27	826	324	0.88	+	0.52	+	10344	4205	0.88	+	0.70	+
72													
73	365	471	286	2.87	–	2.45	–	4879	2620	2.52	–	1.69	–
74	365	825	546	2.68	–	1.55	–	8607	5054	2.27	–	1.26	+
75	365	693	330	3.16	–	2.03	–	7409	3493	3.13	–	1.66	–
76	365	579	335	3.43	–	1.61	+	6317	3638	2.59	–	1.83	–
77	365	620	370	4.16	–	3.50	–	6490	4446	4.54	–	2.83	–
78	272	662	290	2.31	–	1.05	+	6768	2866	2.28	–	1.05	+

[1] k_n and k_{ln} denote the value of the Kolmogorov statistics for normal pdf and log-normal pdf.

[2] Sign + means that at significance level 0.05 there are no reasons to reject the hypothesis about the consistency of empirical distribution with normal or log-normal pdf, respectively.

[3] Sign – denotes that the hypothesis should be rejected.

Relative error

$$\frac{L_R - L_{R+1}}{L_{R+1}} 100 \tag{2}$$

is the *ex post* assessment of the prediction.

It seems that the results contained in tab. 3 cannot be recognized as satisfactory ones. They justify the title of the paper. It should be emphasized that the comparison refers to one value of flow intensity only. For the other ones,

Table 2

Parameters of regression lines for ammonia nitrogen, iron, manganese, and dissolved oxygen versus flow intensities in the Vistula-Warsaw cross-section in the years 1965–1978 in I, II, III models

Parametry linii regresji ładunków azotu amonowego, żelaza, manganu i rozpuszczonego tlenu względem natężenia przepływu w przekroju Wisła–Warszawa w latach 1965–1978 w modelach I, II, III

Year	Model I				Model II				Model III			
	b	Correlation coefficient	k_n		b	Correlation coefficient	k_n		$b \times 10^3$	Correlation coefficient	k_n	
1	2	3	4		5	6	7		8	9	10	
Ammonia nitrogen												
1965	1.07	0.59	2.98	–	1.65	0.16	6.31	–	1.75	0.15	6.19	–
66	0.29	0.68	3.33	–	1.67	0.14	7.95	–	1.69	0.15	7.74	–
67	0.44	0.57	2.93	–	1.19	0.55	1.91	–	1.47	0.55	1.87	–
68	0.97	0.54	2.58	–	1.45	0.41	3.21	–	2.25	0.47	3.08	–
69	1.04	0.49	1.61	–	0.92	0.29	3.04	–	2.24	0.34	2.49	–
70	0.16	0.20	3.71	–	−0.03	−0.01	1.96	–	0.38	0.18	2.12	–
71	0.66	0.56	2.82	–	1.38	0.50	2.19	–	2.29	0.50	2.50	–
72	−0.12	−0.10	3.15	–	−0.82	−0.25	1.45	–	−0.48	−0.10	2.03	–
73	0.05	0.03	3.71	–	0.08	0.03	2.78	–	0.25	0.05	2.74	–
74	0.19	0.17	4.47	–	0.77	0.38	2.10	–	0.75	0.33	2.37	–
75	−0.03	−0.03	3.32	–	−0.02	−0.01	2.91	–		0.06	2.93	–
76	0.39	0.38	3.65	–	1.67	0.53	2.48	–	1.94	0.43	2.88	–
77	0.44	0.38	4.68	–	1.04	0.36	2.15	–	1.37	0.39	1.78	–
78	0.09	0.06	4.10	–	−0.00	−0.00	2.20	–	0.58	0.11	2.46	–
Iron												
1965	15.80	0.69	3.87	–	3.08	0.27	7.18	–	2.98	0.23	6.69	–
66	11.50	0.74	4.50	–	2.83	0.21	7.92	–	2.72	0.20	7.53	–
67	6.84	0.86	3.86	–	2.00	0.91	1.28	+	2.50	0.92	0.70	+
68	7.91	0.78	3.01	–	2.53	0.91	1.20	+	3.44	0.89	1.45	–
69	6.68	0.75	4.32	–	2.13	0.90	1.07	+	4.52	0.91	1.36	–
70	14.56	0.86	3.01	–	2.61	0.92	1.57	–	2.27	0.82	1.94	–
71	5.65	0.83	2.95	–	1.99	0.89	0.70	+	3.41	0.90	1.01	+
72	10.14	0.85	2.19	–	1.98	0.88	0.50	+	2.77	0.87	0.90	+
73	6.43	0.73	4.32	–	1.93	0.88	1.23	+	3.67	0.89	1.23	+
74	5.30	0.76	4.31	–	2.09	0.92	0.68	+	2.18	0.86	0.58	+
75	3.87	0.79	3.44	–	2.04	0.86	0.91	+	2.60	0.87	1.08	+
76	5.87	0.85	3.20	–	2.17	0.91	1.54	–	3.11	0.91	1.12	+
77	3.64	0.86	4.42	–	1.73	0.90	1.11	+	2.10	0.90	1.14	+
78	5.83	0.78	2.88	–	2.22	0.90	0.61	+	3.20	0.92	1.26	+

1	2	3	4		5	6	7		8	9	10	
						Manganese						
1965	0.61	0.76	3.06	−	1.57	0.16	7.13	−	1.61	0.15	6.75	−
66	0.61	0.78	3.83	−	2.14	0.18	8.10	−	2.16	0.19	7.89	−
67	0.33	0.78	3.33	−	1.04	0.82	0.94	+	1.33	0.85	0.96	+
68	0.66	0.78	2.62	−	1.56	0.75	1.40	−	2.33	0.81	0.95	+
69	0.42	0.62	2.05	−	1.14	0.68	1.46	−	2.10	0.64	0.97	+
70	0.98	0.83	2.46	−	1.81	0.54	4.05	−	1.70	0.51	4.27	−
71	0.42	0.71	3.41	−	1.08	0.74	0.93	+	1.92	0.78	0.89	+
72	0.42	0.54	4.16	−	0.80	0.43	1.28	+	1.37	0.52	0.97	+
73	0.28	0.62	3.02	−	0.74	0.49	2.02	−	1.51	0.53	2.03	−
74	0.31	0.70	3.33	−	0.90	0.46	3.25	−	1.09	0.50	3.43	−
75	0.15	0.61	2.91	−	0.78	0.21	5.44	−	1.15	0.24	5.57	−
76	0.39	0.80	2.88	−	1.01	0.75	1.22	+	1.55	0.80	0.68	+
77	0.39	0.82	1.54	−	0.96	0.71	1.33	+	1.26	0.76	1.53	−
78	0.37	0.33	4.55	−	0.55	0.28	1.09	+	1.00	0.37	1.31	+
						Dissolved oxygen						
1967	10.99	0.98	2.09	−	1.00	0.97	1.71	−	1.19	0.94	1.28	+
68	6.13	0.66	1.03	+	0.71	0.73	0.93	+	0.95	0.69	1.04	+
69												
70	8.78	0.93	0.64	+	1.14	0.84	1.66	−	0.87	0.69	2.13	−
71	12.62	0.97	1.09	+	0.98	0.96	1.27	+	1.09	0.95	0.58	+
72												
73	8.39	0.92	2.23	−	0.93	0.96	1.48	−	1.59	0.88	1.12	+
74	8.76	0.95	1.70	−	0.90	0.97	1.07	+	0.92	0.89	1.11	+
75	9.81	0.93	2.25	−	0.88	0.93	1.28	+	1.09	0.91	0.94	+
76	10.44	0.96	1.02	+	1.03	0.96	1.36	−	1.38	0.90	1.68	−
77	11.68	0.97	1.64	−	1.07	0.94	1.20	+	1.27	0.92	1.55	−
78	9.25	0.94	1.31	+	0.89	0.93	1.32	+	1.23	0.91	1.11	+

k_n, + and − have the same meanings as in tab. 1 related to residual variables. Natural logarithm has been used.

differing more than 20% from the mean value, the errors would be more frequently greater.

5. The author is rather sceptical about the use of other regression models (with greater number of parameters or of other shapes) as prediction models for forecasting water quality when a set of data (Q, L), formally treated, becomes the basis for model construction. The information about expected changes of the economical state in the river basin and expected modifications of production engineering seems to be indispensable.

6. The research for this paper was done at the Institute of Meteorology and Water Economy, Wrocław. The author is much indebted to Mr. M. Czapliński and Mrs. M. Kapłańska for writing computer programmes and their assistance in completing the outputs from printouts.

Table 3

Relative estimate *ex post* of load predictions for ammonia nitrogen, iron, manganese, and dissolved oxygen which would flow through the Vistula–Warsaw cross-section in the year $R+1$, if the flow intensities were equal to 1.20_R where Q_R is the mean flow intensity in the year R. Relative error of the prediction is computed from formula (2)

Względna ocena *ex post* prognoz ładunków azotu amonowego, żelaza, manganu i rozpuszczonego tlenu, jakie przepłynęłyby w roku $R+1$ przez przekrój Wisła–Warszawa, gdyby pojawiły się natężenia przepływu 1.20_R, gdzie Q_R jest średnim natężeniem przepływu w roku R. Błąd względny prognozy jest obliczony ze wzoru (2)

Year	Model					
	I	II	III	I	II	III
	Relative prediction error, %.					
	Ammonia nitrogen			Iron		
1966	188	38	30	12	11	−13
67	−55	−94	−94	22	−89	−90
68	−34	−18	−37	30	−26	−43
69	−35	−17	−32	−51	−18	−37
70	55	6.2	5.7		65	27
71	−31	49	−41	63	29	−21
72	44	105	43	−31	−11	−7.5
73	−27	−34	−23	10	7.9	9.0
74	13	−89	26	53	−85	27
75	30	29	2.8	26	5.0	−17
76	−23	−86	−2.1	40	24	−20
77	−73	12	−7.8	6.0	4.4	−2.0
78	30	108	64	70	12	−20
	Manganese			Dissolved oxygen		
1966	18	30	29			
67	12	−90	−90			
68	−28	−91	−32	26	20	1.3
69	−19	−29	−28			
70	587	135	89			
71	43	7.5	−18	−18	19	−28
72	−2.4	7.5	2.2			
73	10	3.8	5.5			
74	102	27	26	−10	6.1	−15
75	44	35	12	−2.6	0.03	−6.8
76	−53	−54	−51	−2.1	0.79	−1.3
77	−8.8	−19	−16	2.2	4.8	6.2
78	−16	9.3	−7.3	5.5	1.0	−6.6

CZY METODA REGRESYJNA JEST DOBRYM NARZĘDZIEM PROGNOZOWANIA JAKOŚCI WÓD?

Celem pracy jest ocena weryfikalności regresyjnego prognozowania ładunku zanieczyszczeń (rozpuszczonego tlenu), a zatem i stężeń przepływających przez pewien przekrój rzeki. Weryfikowane są trzy typy modeli: liniowy, logarytmiczno–logarytmiczny i logarytmiczno–liniowy. Weryfikacja obejmuje dane z przekroju Warszawa–Wisła : (i) azot amonowy, żelazo i mangan, 14 lat, (ii) tlen rozpuszczony, 10 lat. Weryfikowane są hipotezy zgodności rozkładów empirycznych ładunków i natężeń przepływów z odpowiednimi rozkładami teoretycznymi (test Kołmogorowa). Obliczono dla każdego roku parametry linii regresji oraz ładunki zdefiniowane liniami regresji dla pewnego wybranego natężenia przepływu. Wyniki weryfikacji są zebrane w trzech tablicach. Śledząc je można odkryć niesatysfakcjonującą niestabilność, z roku na rok, parametrów linii regresji i ładunków przepływających przez przekrój rzeki, gdy pojawiają się te same intensywności przepływu.

IST DIE REGRESSION EIN GUTES INSTRUMENT ZUR VORHERSAGE DER WASSERGÜTE?

Das Ziel der Untersuchung ist die Verifizierbarkeit einer regression-gestützten Vorhersage der Belastung (gelöster Sauerstoff) – und damit auch der Konzentration der durchfließenden Verunreinigungen – für einen gewissen Flußquerschnitt. Verifiziert werden die folgenden drei Modelle: das lineare, das logarithmisch–logarithmische und das logarithmisch–lineare Modell. Die Verifikation bezieht sich auf die Daten für den Querschnitt Warschau-Weichsel und umfaßt: (i) Ammoniumstickstoff, Eisen und Mangan für den Zeitraum von 14 Jahren und (ii) gelösten Sauerstoff für den Zeitraum von 10 Jahren. Es werden auch die Hypothesen der Konkordanz empirischer Verteilungen von Belastung- und Durchflußstärke mit theoretischen Verteilungen (Kolmogorov-Test) verifiziert. Für jedes Jahr werden die Parameter der Regressionslinien, sowie die mit Hilfe der Regressionslinien für eine gewisse Durchflußstärke bestimmten Belastungen berechnet. Die Ergebnisse der Verifikation werden in drei Tabellen zusammengestellt. Ihre Analyse kann zu der Entdeckung einer nicht befriedigenden Unbeständigkeit von den Parametern der Regressionslinien und auch der den Flußquerschnitt durchfließenden Belastungen führen, wenn dieselbe Durchflußstärke auftritt.

ЯВЛЯЕТСЯ ЛИ РЕГРЕССИОННЫЙ МЕТОД ХОРОШИМ ИНСТРУМЕНТОМ ПРОГНОЗИРОВАНИЯ КАЧЕСТВА ВОДЫ?

Целью работы является оценка верифицируемости регрессионного прогнозирования количества загрязнений (растворимого кислорода), а тем самым и концентраций, протекающих через некоторый профиль реки. Верифицируются три типа моделей: линейная, логарифмо–логарифимическая и логарифмо–линейная. Верификация охватывает данные из сечения Варшава-Висла: (i) аммиачный азот, железо и магний, 14 лет, (ii) растворенный кислород, 10 лет. Верифицируются гипотезы соответствия эмпирических количеств и расходов с соответствующими теоретическими распределениями (критерий Колмогорова). Были рассчитаны для каждого года параметры линии регрессии, а также количества, определённые линиями регрессии для некоторого избранного расхода. Результаты верификации собраны в трёх таблицах. Следя за ними, можно выявить неудовлетворяющую нестабильность, из года в год, параметров линии регрессии и количеств загрязнения, протекающих через профиль реки, когда появляются те же расходы.