Issues of accounting for outliers in assessing of the nutrients runoff

Vladislav Shelutko* and *Maria* Makarova

Russian State Hydrometeorological University, 79 Voronezhskaya ul., 192007, St. Petersburg, Russia

Abstract. The research work is devoted to the analysis of the highest values of nutrient concentrations in the runoff in the Velikaya River for the period 1969–2009. According to the results of the study, when assessing the numerical characteristics of river water pollution, it is necessary to exclude outliers from the observation series. The presence of outliers in the calculation together with the main observational data leads to a significant overestimation of the numerical characteristics of river pollution in the average annual and multi-year period. In this case, it is necessary to study the characteristics of the outliers themselves, regardless of the initial samples. The paper presents an attempt to define the probability of outliers by combining outliers information on individual observation series.

1 Introduction

Usually, in the studies of river pollution, the scientists bring the most attention to the analysis of averaged values of indicators of the ecological state of the environment. In this case, possible deviations of the indicator values from the average value are either not taken into account at all or are not taken into account sufficiently[1]. Meanwhile, sometimes, these deviations determine the possible extreme values and should be taken into account in solving practical environmental tasks.

Thus, the study of the methods for identifying and assessing the significance of extreme values or the so-called “outliers” [1] is an important task in the practice of statistical analysis. In this regard, the main goal of the research is to analyze the outliers in the series of nutrient concentrations in the Velikaya river.

2 Materials and methods

The source of the Velikaya River is located near the village of Shepeli, it flows into the Pskov Lake, 4 km west of the village of Murovitsy. The river is 430 km long, the catchment area is 25,200 km^2 [2]. Three stationary observation points are installed on the Velikaya River. They located in proximity to the towns of Opochka, Ostrov, and Pskov. Each observation point has two sections – above (upstream cross-section) and below (downstream cross-section) the city line. The observation period under consideration is 1969 – 2009.

We used long-term observation data on the concentrations of the following elements:

*Corresponding author: shelutko@rshu.ru

nitrate-nitrogen ($N\text{-}NO_3^-$), ammonium nitrogen ($N\text{-}NH_4^+$), total iron (Fe_{total}), and the indicator of biological oxygen demand for 5 days (BOD_5). The data was provided by the Northwest Department of Hydrometeorology and Environmental Monitoring.

The study used the methods of preliminary statistical analysis of data, including a descriptive analysis of time series, parametric methods, as well as visual and calculated methods of estimating outliers.

3 Results and discussions

The first stage of the study is the analysis of the distribution parameters of the initial data series. Table 1 shows the estimates of numerical characteristics for all initial data series.

Table 1. Estimates of the numerical characteristics of nutrient concentrations in water (mg/dm^3) from the initial observation series: *m*–mean values, σ– standard deviation, C_v– coefficient of variation, g– skewness coefficient.

Observation points	Nutrient	Section	Numerical characteristics					MPC mg/dm^3
			m	**σ**	**C_v**	**g**	**g/Cv**	
Opochka	$N\text{-}NH_4^+$	Upstream	0.11	0.15	1.36	1.87	1.38	0.4
		Downstream	0.12	0.15	1.26	1.63	1.29	
	BOD_5	Upstream	1.75	0.83	0.48	0.51	1.06	2.1
		Downstream	2.13	1.32	0.62	3.61	5.82	
	Fe_{total}	Upstream	0.10	0.10	1.00	1.88	1.88	0.1
		Downstream	0.10	0.09	0.89	1.21	1.36	
	$N\text{-}NO_3^-$	Upstream	0.32	0.29	0.90	1.59	1.77	9
		Downstream	0.38	0.33	0.87	1.32	1.52	
Ostrov	$N\text{-}NH_4^+$	Upstream	0.17	0.28	1.65	2.65	1.61	0.4
		Downstream	0.19	0.28	1.50	2.02	1.35	
	BOD_5	Upstream	1.57	0.69	0.44	1.03	2.34	2.1
		Downstream	1.93	0.99	0.51	1.55	3.04	
	Fe_{total}	Upstream	0.17	0.19	1.10	3.14	2.85	0.1
		Downstream	0.16	0.15	0.97	1.34	1.38	
	$N\text{-}NO_3^-$	Upstream	0.39	0.33	0.86	1.12	1.30	9
		Downstream	0.41	0.40	0.98	2.14	2.18	
Pskov	$N\text{-}NH_4^+$	Upstream	0.22	0.30	1.38	1.81	1.31	0.4
		Downstream	0.24	0.31	1.27	1.51	1.19	
	BOD_5	Upstream	1.89	0.99	0.52	2.21	4.25	2.1
		Downstream	2.09	1.23	0.59	2.70	4.58	
	Fe_{total}	Upstream	0.19	0.17	0.92	1.20	1.30	0.1
		Downstream	0.20	0.20	0.98	1.59	1.62	
	$N\text{-}NO_3^-$	Upstream	0.49	0.45	0.92	1.50	1.63	9
		Downstream	0.53	0.62	1.17	4.85	4.14	

According to the calculation results presented in Table 1, one may note some features of the concentration distribution along the length of the Velikaya River. For example, the main increase in the values of the skewness coefficient (*g)* is the most often observed within the urban area. Moreover, if the estimates of the numerical characteristics *m* and C_v vary over the sections and elements within small limits, the values of *g* from section to section often increase more than three times. Such big changes in the values of C_s with small changes in values of *m* and C_v the most often happen in cases where there are extreme values in the series of observations that much deviate from the values of all other members of the series – the so-called outliers. By definition, the outliers are the observations considerably higher or lower than most of the data, which infrequently but regularly occur [3].

As noted earlier [4], the outliers are caused by enterprises' emergency discharges or extreme weather. So, the genesis of outlier's formation differs from the main observational data.

To confirm the presence of outliers in the initial data series it is necessary to find the distribution laws from that series. At the same time, the standard distribution law, the law of Pearson type III, Kritsky-Menkel law, the log-normal law, and the Gambel distribution law are considered as the famous theoretical laws in statistical calculations.

Verification of the coincidence of empirical probability curves (the coordinates calculated directly from the initial series of observations) and theoretical probability curves according to the test Cramer-Mises-Smirnov $n\omega^2$ are carried out. According to [5], the value of the upper confidence limit of the statistics $n\omega^2$ at a 5% significance level is 0.461.

According to the results the distribution laws of Pearson type III (88% of the cases considered) and Kritsky-Menkel (12% of cases) is optimal in this case.

A visual analysis of the coincidence of the selected theoretical and empirical probability curves showed that there are empirical points in some series which are considerably higher than most of the data. For all series of observations, 28 such values were discovered. These concentration values deviating from the probability curves need to be taken as outliers. Verification of this assumption was carried out by testing statistical hypotheses according to the criteria of Dixon and Smirnov-Grabbs.

As the analysis showed for 14 values out of 28, the hypothesis refuted, which indicates that they do not belong to the initial series. This means that these values are the outliers. Thus, 4 outliers detected in the series of nitrate-nitrogen ($N\text{-}NO_3^-$), 2 outliers in the series of total iron (Fe_{total}), and 8 outliers in the series of BOD_5.

Thus, two groups of observation data are selected in each initial series containing anomalous values. The first group dictates the choice of the theoretical distribution law and reflects the regular regime of watercourse pollution permissible in the conditions of the planned operation of enterprises. It comprises most of the values of the initial series. The second group (outliers) is not a subject to the estimated distribution laws and most likely reflects the pollution regime in emergency discharges or extreme weather [1].

The outliers, in most cases, last no more than a few days. At the same time, when average annual concentrations or volumes of pollutant runoff are calculated, the same weight given to each measurement, including outliers. For example, with conducted 12 measurements of concentration per year, each measurement in determining the average annual concentration is given a weight of 30 days. This leads to an unjustified overestimation of the numerical characteristics of the concentrations. In this regard, the concentrations attributed to the outliers are excluded from the calculation of the numerical characteristics of the series (Table 2).

As follows from the presented data, by excluding outliers from the calculation, the average annual concentrations decreased by 5–10%. The measures of spread and skewness changed even more significantly. For example, the coefficient of variation (C_v) decreased by 10–20% and the skewness coefficient (g) decreased in some cases more than three times. Regarding this, it is possible to significantly reduce the extreme values of the nutrient concentrations during the normal planned operation of enterprises. For example, the possible values of nutrient concentrations with a probability of 1% (C_1%) decrease on average by 5–20%, if the outliers are excluded. The inclusion of outliers of data beyond the scope of these samples in the general statistical analysis can significantly distort the results of the study and lead to an artificial overestimation of the characteristics of nutrient runoff.

On the other hand, as noted earlier, anomalous values (outliers) characterize the pollution regime of the watercourse during possible emergencies or extreme weather conditions. It is interesting and practically important to assess the extent of pollutants' concentration variability before described conditions.

Table 2. Comparison of estimates of numerical characteristics of nutrient concentrations in water according to the initial observation series and when excluding outliers.

Observation points	Nutrient	Section	Numerical characteristics							
			m	$C_{1\%}$	C_v	g	*m*	$C_{1\%}$	C_v	g
			Initial observation series				When excluding outliers			
Opochka	BOD_5	Downstream	2.13	6.32	0.63	3.56	2.05	4.99	0.48	0.27
Ostrov	Fe_{total}	Upstream	0.17	0.85	1.10	3.17	0.16	0.69	0.94	1.32
	$N\text{-}NO_3^-$	Downstream	0.41	1.83	0.98	2.17	0.38	1.48	0.85	1.02
Pskov	BOD_5	Upstream	1.89	4.87	0.52	2.22	1.84	4.34	0.46	1.00
		Downstream	2.09	5.90	0.59	2.71	2.01	4.91	0.48	1.00
	Fe_{total}	Downstream	0.20	0.91	0.98	1.60	0.20	0.86	0.94	1.25
	$N\text{-}NO_3^-$	Downstream	0.53	2.82	1.18	4.89	0.48	1.83	0.83	1.01

To assess the possible values of outliers it is necessary to find the law of their distribution. The determination of the theoretical curve of outliers for each series of observations is impossible in this case due to the lack of sufficient initial empirical material. In this regard, a decision was made, which based on combining all outliers for different elements and different data series into one sample. Moreover, it is assumed that the normalized initial series and time series of outliers are stationary and ergodic. The need to take into account these features of the initial series of observations of the hydrochemical regime of rivers is theoretically justified [6, 7].

The combined values of outliers relate to different series of observations, and each of them has own numerical characteristics. Therefore, the concentration values are necessary to be normalized according to the mean value and standard deviation.

$$t_{ji} = \frac{x_{ji} - m_j}{\sigma_j} \tag{1}$$

Where j is the number of the series of observations, i is the number of the outlier in this data series, x_{ji} is the concentration of the ith outliers in the jth data series, m_j and σ_j are the mean value and standard deviation of the jth data series, respectively. The values of the numerical characteristics of the formed series are shown in Table 3.

Table 3. Estimates of the numerical characteristics of normalized outliers for the combined series.

Numerical characteristics				
m	s^2	σ	C_v	g
5.82	2.56	1.599	0.275	0.553

As the results of the assessment showed, the type of optimal theoretical probability curves is identical to the type used for the initial series. Therefore, the Pearson type III distribution law can be considered optimal for the combined sample. Table 4 shows the results of calculating the coordinates of the probability curves for the combined series of outliers in normalized ordinates (t_p) and modular coefficients (K_p).

Table 4. The coordinates of the probability curves of the combined series of outliers of nutrient concentrations.

P %	0.01	0.1	1	5	10	25	50	90	95	99	99.9
t_p	5.55	4.54	2.90	1.84	1.34	0.58	–0.13	–1.17	–1.37	–1.73	–2.00
K_p	2.52	2.25	1.80	1.51	1.37	1.16	0.96	0.68	0.62	0.53	0.45

The obtained coordinates of the optimal distribution law show the normalized values of

outliers for all considered data series. To move from normalized values (t_p) to real values for a particular element and observation point it is necessary to convert the values of normalized ordinates of the optimal probability curves to natural concentrations:

$$x_{jp} = m_j + \sigma_j t_{jp}, \tag{2}$$

where x_{jp} is the value of the concentration of outliers of a given probability P, m_j is the mean value, σ_j is the standard deviation, t_{jp} is the normalized ordinate of the given probability. The results of such a calculation are shown in Table 5.

Table 5. Values of outliers of concentrations of nutrients and BOD_5 of various probability(mg/dm^3).

Observation points	Nutrient	Section	P, %							
			0.1	**1**	**5**	**10**	**25**	**50**	**90**	**99.9**
Opochka	BOD_5	Downstream	4.78	3.83	3.21	2.91	2.47	2.05	1.45	0.96
Ostrov	Fe_{total}	Upstream	0.38	0.31	0.26	0.23	0.2	0.16	0.12	0.08
	$N\text{-}NO_3^-$	Downstream	0.92	0.74	0.62	0.56	0.47	0.39	0.28	0.18
Pskov	BOD_5	Upstream	4.24	3.39	2.84	2.58	2.19	1.82	1.28	0.85
		Downstream	4.7	3.76	3.15	2.86	2.43	2.02	1.42	0.95
	Fe_{total}	Downstream	0.46	0.37	0.31	0.28	0.24	0.2	0.14	0.09
	$N\text{-}NO_3^-$	Downstream	1.18	0.94	0.79	0.72	0.61	0.51	0.36	0.24

These data are the values of concentrations that a given random variable would exceed with a fixed probability if the outlier occurred. For example, an excess of nitrate-nitrogen concentration at the Ostrov observation point (downstream cross-section) of 0.92 mg/dm^3 can occur with a probability of 0.1%. Thus, this case is much unlikely.

To calculate the exceeding probability at a particular time it is necessary to take into accounts the probability of the outliers themselves:

$$P_\beta = \frac{m_i}{n_i} \cdot 100\%, \tag{3}$$

where m_i and n_i are the number of outliers and the total number of observations in the ith observation series, respectively.

Thus, the probability of exceeding defined as the product (P_α) and the outliers' frequency of the ith element in the river (P_β):

$$P_i = P_\alpha \cdot P_\beta, \tag{4}$$

According to the results of calculating the frequency of outliers, the probability of occurrence of anomalous values in a series of observations of the BOD_5 indicator within the Opochka observation point item was 0.78%; total iron (Fe_{total}) and nitrate-nitrogen ($N\text{-}NO_3^-$) within the Ostrov observation point – 0.68 and 1.66%, respectively; BOD_5 in the upstream and downstream cross-section, total iron (Fe_{total}) and nitrate-nitrogen ($N\text{-}NO_3^-$) within the Pskov observation point - 0.90, 1.23, 0.63, and 1.03%, respectively.

So, such as taking into account the optimal distribution law and the provisions described above, it turns out that within the Ostrov observation point, the nitrate-nitrogen concentration values ($N\text{-}NO_3^-$) in water with a probability of 0,1 and 99% can be greater than 1.53 mg/dm^3 or equal to 0.37 mg/dm^3. Thus, the extremes of the concentrations of rare and most frequent repeatability were determined. Indeed, the greatest concentration value in this sample illuminated by the observation period is 1.38 mg/dm^3, and the sample average is 0.38 mg/dm^3.

4 Conclusion

As the results of the study show, the outliers are observed in some data series. In this case, the outliers are the values of concentrations of pollutants significantly exceeding the concentration of most values of observations data. The reason for their presence in the series of observations may be emergency discharges of enterprises or extreme weather. The inclusion of outliers in the calculation along with the main observational data leads to a significant overestimation of the numerical characteristics of river pollution in the average annual and multi-year period.

Therefore, when assessing water pollution levels under normal conditions, it is necessary to estimate outliers at the preliminary analysis stage and exclude them from calculations.

However, it is necessary to assess the probability and possible values of the outliers themselves. In this case, proposed outliers assessment method is to combine the normalized outliers values for different series of observations in the river system into one series. The basis of this method is the assumption that the processes under consideration are stationary and ergodic. Therefore, it gives approximate characteristics of outliers.

This method for estimating outliers allows making significant adjustments in the calculation of nutrient transfer with river runoff at the preliminary analysis stage.

Authors would like to thank the Northwest Department of Hydrometeorology and Environmental Monitoring for making the data used in this article available to us.

References

1. V. Shelutko, Issues of applied ecology. Collection of scientific papers, 15-23 (2002) (in Russian)
2. Schemes of Integrated Use of Water Bodies of the Narva River Basin, **1,** 17 (2014) (in Russian)
3. D.R. Helsel, R.M. Hirsch. *Statistical methods in water resources* (U.S. Geological Survey, Reston, 2002)
4. V. Shelutko, E. Smyzhova, Proceedings of the RSHU,**13**, 89-104 (2010) (in Russian)
5. R 50.1.037–2002. Applied statistics. Rules of check of experimental and theoretical distribution of the consent. Part II. Non parametric goodness-of-fittest (in Russian)
6. V. Shelutko, E. Smyzhova, Materials of the international scientific-practical conference "Geographical education and science in Russia: history and current state", 862-871 (2010) (in Russian)
7. E. Urusova, The meteorological bulletin, **9(2)**, 216-220 (2017) (in Russian)