JOURNAL OF WATER AND LAND DEVELOPMENT
e-ISSN 2083-4535

Polish Academy of Sciences (PAN) Institute of Technology and Life Sciences – National Research Institute (ITP – PIB)

JOURNAL OF WATER AND LAND DEVELOPMENT
DOI: 10.24425/jwld.2022.142313
2022, No. 55 (X–XII): 115–124

Analysis of the distribution of statistical concentrations of pollutants in samples of treated wastewater from small sewage treatment plants

Grzegorz B. Kaczor

University of Agriculture in Krakow, Faculty of Environmental Engineering and Land Surveying,
Al. Mickiewicza 24/28, 30-059 Kraków, Poland

RECEIVED 12.04.2022 ACCEPTED 09.06.2022 AVAILABLE ONLINE 19.12.2022

Abstract: The aim of the research was to show which theoretical statistical distribution best reflects and describes the variability of pollutant concentrations in treated sewage, discharged from small sewage treatment plants, characterised by a value below 2000 PE. The statistical analysis additionally takes into account the influence of the number of measuring sequence data on the shape and level of the distribution fit. The data for the research were obtained from three small sewage treatment plants, operating in the Lesser Poland, 10, 11 and 14 km from Kraków. Due to their size, these facilities are included in the group of treatment plants below 2000 PE. The research was conducted for 10 years. In the statistical analysis, 20-, 40-, 60- and 80-element data series were used, including the values of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD_{Cr}) and total suspended solids (TSS), determined in samples of treated wastewater. Two commonly used tests, Kolmogorov–Smirnov λ and Pearson's χ^2 test were used to assess the fit of the theoretical statistical distribution to the empirical data distribution. Statistical analysis showed that the studied communities were characterised by an asymmetric, right-oblique distribution. Most often, the empirical distribution of the analysed measurement sequences was consistent with the Fisher–Tippett distribution. On the basis of the χ^2 test, this distribution was described by a total of 31 out of 36 analysed groups at the significance level of $a = 0.05$. Other distributions that often describe the analysed empirical data are: Gamma, log-normal, Chi-square, and Weibull. The common feature of these distributions is usually asymmetry, right oblique. The skewness value ranges from 0.15 to 1.69.

Keywords: pollutants, sewer system, significance test, statistical distributions, treatment plant

INTRODUCTION

In recent years, there has been an increased interest in issues related to risk and failure rate analysis, hazard identification and the reliability of operation of water supply systems as well as sewage disposal and treatment systems [Andraka 2011; Meijer *et al.* 2018; Młyński *et al.* 2016a; 2020; Oliveira, von Spearling 2008; Taheriyoun, Moradinejad 2015; van Riel *et al.* 2015]. In Poland, it is related to the introduction by the new Water Law of the obligation to analyse and assess risk in water intake and distribution systems. Currently, risk assessment and analysis are also implemented in sewer networks and wastewater treatment plants. In the case of large water supply and sewage systems, operating within the boundaries of urban agglomerations, many procedures and analytical methods regarding risk analysis and reliability of their operation have already been implemented [Andraka, Dzienis 2003; Meijer *et al.* 2018; Oliveira, von Spearling 2008; Śliz 2018; Tchórzewska-Cieślak, Piegdoń 2016]. However, there is still insufficient action in this regard in the case of small sewage systems [Bugajski *et al.* 2017; Marzec 2017; Nastawny, Jucherski 2013; Wałęga 2009].

Most often, in scientific works related to the assessment of the reliability of the operation of sewage treatment plants, the values of pollution indicators in treated wastewater samples are compared with the limit values specified in the water-legal permit or the Regulation of the Minister of Maritime Economy and

Inland Navigation of 12 July 2019 on substances particularly harmful to the aquatic environment and the conditions [Rozporządzenie … 2019] to be met when discharging sewage into waters or ground, as well as when discharging rainwater or snowmelt into waters or into water facilities [Chmielowski *et al.* 2009; 2015; Młyńska *et al.* 2017; Młyński *et al.* 2016b].

At the same time, in some works, the authors try to use advanced statistical methods to assess the operation of wastewater treatment plants, allowing for the extension of the information obtained on the duration of failures, process stability and the forecast of the reliability of neutralisation of individual pollutants [Andraka 2005; 2011; Młyński *et al.* 2016; 2020; Oliveira *et al.* 2012; Oliveira, von Spearling 2008; Siwiec *et al.* 2018; Wałęga 2009; Zawadzka *et al.* 2021]. Using advanced statistical methods in calculations and reliability analyses, it is assumed that the values of the analysed pollution indicators in raw and treated sewage are subject to variability according to a random function. In such analyses, it is necessary to know the statistical distribution of these data. It is surprising that individual researchers point to different theoretical statistical distributions as best describing the variability in the quality of treated wastewater. Niku *et al.* [1981], Andraka [2005; 2011; 2020] and Oliveira *et al.* [2012] in their statistical calculations of the reliability of sewage treatment plant operation indicated and used the log-normal distribution as the best representation of the variables. Bugajski *et al.* [2012]; Bugajski [2014a, b]; Bugajski and Nowobilska-Majewska [2019], Nastawny and Jucherski [2013], Wąsik *et al.* [2016], Marzec [2017], and Zawadzka *et al.* [2021] indicated the Weibull distribution. Other distributions have been suggested in the works of Młyński *et al.* [2016a; 2020].

Taking into account the discrepancies between the types of adopted distributions and the results of their matching, tests and analyses were carried out to show which theoretical statistical distributions best describe the variability of pollutant concentrations in treated wastewater discharged from small wastewater treatment plants below 2000 PE. Additionally, the influence of the size of the measurement sequence data on the shape and level of adjustment of the distribution was taken into account.

MATERIALS AND METHODS

The data for the research were obtained from three small sewage treatment plants, located in Lesser Poland, 10 (A), 11 (B) and 14 km (C) from Kraków. Due to their size, these facilities are included in the group of treatment plants below 2000 PE. Their general characteristics are presented in Table 1.

These treatment plants discharge treated wastewater into flowing waters, therefore the permissible values for total nitrogen and total phosphorus are not specified in the water-legal permits. Therefore, only the values of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD_{cr}) and total suspended solids (*TSS*) were included in the statistical analysis. The quality of raw and treated sewage was tested for 10 years (2005–2015). During this period, 80 samples of treated wastewater were collected at each facility (on average 8 samples per year). The treated wastewater was tested in the same accredited laboratory in Krakow. In the case of small wastewater treatment plants, up to 2000 PE, the formal and legal conditions require the collection of at least 4 wastewater samples per year [Rozporządzenie … 2019]. The study assumed the collection of 8 samples per year. In the statistical analyses, 40 samples more were used than was possible based on the archival data.

Table 1. Basic characteristics of analysed wastewater treatment plants

Parameter	Value of the parameter for a given treatment plant		
	treatment plant A	treatment plant B	treatment plant C
Size of treatment plant by PE	1530	1280	1960
Average daily sewage inflow ($m^3 \cdot d^{-1}$)	239.2	224.0	253.9
Technological system of the sewage treatment plant	Huber screen, Imhoff primary settling tank, flow reactor with nitrification and denitrification chamber, vertical secondary settling tank		
Type of wastewater	domestic from housing and public facility		
Type of sewage system	separate gravity, made of stoneware		
Number of sewage samples taken	80	80	80

Source: own study.

BOD_5 was determined by the dilution and grafting method with the addition of allylthiourea [PN-EN ISO 5815-1:2019-12], the *COD* value was determined by the bichromate method [PN-ISO 6060: 2006], and the total suspended solids content by filtration through glass fiber filters [PN-EN 872:2007].

With regard to the conditions of the water-legal permits of the analysed sewage treatment plants, the BOD_5 values in treated sewage may not exceed 35 $mg \cdot dm^{-3}$, *COD* – 125 $mg \cdot dm^{-3}$, and the total suspended solids – 35 $mg \cdot dm^{-3}$.

In the studies for each of the three treatment plants, 80-element data observation sequences were used, including the values of BOD_5, *COD* and total suspended solids determined in the treated wastewater samples. Data analysis in terms of fitting the statistical distribution was performed sequentially for the first 20, 40, 60 and 80 samples. Such a division showed the influence of the number of samples on the shape and fit of the theoretical distribution. It was considered that the data in individual groups of 20, 40, 60 and 80 elements will be summarised the most representative if they are divided only according to the date of sewage sampling. Therefore, the first group, 20 elements, was created from the data from 1st Jan 2005 to 30th Jun 2007. A group of 40 elements from the data from 1st Jan 2005 to 31st Dec 2009. The 60-element group from the data from 1st Jan 2005 to 30th Jun 2012. The 80-element group from the data from 1st Jan 2005 to 31st Dec 2014. This grouping criterion made it possible to maintain seasonal relationships in individual groups and not to mix individual data with each other.

In order to obtain reliable and compared results of the selection of the statistical distribution, the number of class intervals k was calculated for each number of the measuring sequence according to the Equation (1) [Luszniewicz, Słaby 2003; Sobczyk 2022]:

$$k = \sqrt{n} \quad (1)$$

where: k = number of class intervals, n = number of elements in the sample or measurement sequence.

Based on Equation (1), 5 class intervals were adopted for 20-element series, 6 for 40, 8 for 60, and 9 for 80.

Two commonly used tests, the Kolmogorov–Smirnow λ test and the Pearson χ^2 test were used to assess the fit of the theoretical statistical distribution to the empirical distribution.

The Kolmogorov–Smirnow test uses the supremum distance between the empirical distribution function $F(X)$ and the theoretical distribution factor $F_0(X)$. With a random sample of the values of pollution indicators X_1, X_2, ..., X_n, coming from the distribution with unknown distribution factor F, the hypothesis was tested:

$$H_0 : F(X) = F_0(X) \quad (2)$$

stating that the distribution F for all $X \in (-\infty; \infty)$ is equal to a certain determined distribution $F_0(X)$. The alternative hypothesis was considered to be:

$$H_1 : F(X) \neq F_0(X) \quad (3)$$

The Kolmogorov test statistic (D_n) is:

$$D_n = \max_{-\infty < X < \infty} |F(X) - F_0(X)| \quad (4)$$

Based on the D_n statistic, the λ statistic expressed by the Equation (5) was determined:

$$\lambda = D_n \sqrt{n} \quad (5)$$

The concordance test χ^2 is the most frequently used non-parametric test, used to verify the hypothesis H_0 that the observed feature X, in the general population, has a specific type of distribution. With a random sample of the values of pollution indicators X_1, X_2, ..., X_n, derived from the distribution with unknown distribution factor F, the H_0 hypothesis was tested according to Equation (2) in opposition to the alternative H_1 according to Equation (3):

$$H_0 : F(X) = F_0(X)$$

$$H_1 : F(X) \neq F_0(X)$$

The statistic has a distribution χ^2 with k – 1 degrees of freedom (Eq. 6):

$$X_c^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \quad (6)$$

where: c = number of degrees of freedom, k = number of elements, n_i = empirical value, np_i = theoretical (expected) value resulting from the hypothesis corresponding to the measured value.

From the distribution tables χ^2 the critical value χ^2 is read.

The initial analysis of the distributions in conjunction with the results of the research of other authors [Andraka 2005; 2011; 2020; Bugajski *et al.* 2012; Bugajski 2014a, b; Bugajski, Nowobilska-Majewska 2019; Marzec 2017; Niku *et al.* 1981; Oliveira *et al.* 2012; Wąsik *et al.* 2016] decided to try to fit such theoretical distributions as: Chi-square, Erlang, Fisher–Tippett, Gamma, General Extreme Values distribution (GEV), log-normal, logistic, normal and Weibull.

The level of adjustment to the distribution of the values of impurity indicators of theoretical distributions was tested using both methods at the significance level of a = 0.05. Distribution parameters were estimated using the maximum likelihood method. In the study, nine theoretical distributions were adjusted to 4 data groups, separately for each of the 3 pollution indicators, separately for 3 research objects. A total of 324 distribution matches were made, and each fit was assessed with 2 significance tests.

RESULTS AND DISCUSSION

The statistical analysis used the values of BOD_5, COD_{Cr} and total suspended solids in samples of treated wastewater, in accordance with the methodology. Table 2 presents the statistical characteristics of the analysed data. This ruled out the appearance of extreme values (treated as statistically outliers) that disturb the general nature of the data distribution and worsen the degree of its fit. Preliminary box-and-whisker plot and additionally Grubbs test were used to reject outliers.

The obtained results indicate a very similar composition of treated wastewater discharged to the receiving body from all three tested sewage treatment plants (Tab. 2). The average BOD_5 values in the three analysed sites differ by a maximum of 2.8 mg·dm^{-3} (i.e. by 22.2%), COD_{Cr} – by 11.1 mg·dm^{-3} (18.9%), and the total suspended solids – by 1.7 mg·dm^{-3} (10.3%). The values of the standard deviation of BOD_5 do not differ by more than 2.1 mg·dm^{-3}, COD_{Cr} – 3.4 mg·dm^{-3}, and suspensions – 0.4 mg·dm^{-3}. Raw data show asymmetry of distribution, because in the case of BOD_5 the mean is higher than the median by a maximum of 1.7 mg·dm^{-3}, COD_{Cr} – by 2.3 mg·dm^{-3}, total suspended solids – by 1.1 mg·dm^{-3}. The minimum and maximum values indicate a similar range of the analysed data. Preliminary data analysis shows a similar differentiation and range of variability of the analysed pollutants. Therefore, the question arises whether the demonstrated similar variability of data will result in their similar theoretical distributions.

During the statistical analysis, the null hypothesis of H_0 was verified; that the value of a given pollutant index is subject to a specific theoretical decomposition. The alternative hypothesis H_1 states that the empirical distribution of a given variable is not consistent with the distribution adopted in the H_0 hypothesis. Tables 3–5 present the p-values of the probability of not rejecting the null hypothesis about the fit of a given distribution using the Kolmogorov–Smirnov concordance test and the χ^2 concordance test. The measurement data of a given pollution index was divided into 20, 40, 60 and 80 element sequences in order to show how the amount of data influences the fit and shape of the theoretical distribution. In other works, the Authors based their statistical analyses on the following number of measurement sequences: Bugajski *et al.* [2012] – 2 years of research – 36-element sequences, Nastawny and Jucherski [2013] – 8 years of research – 53-element sequences, Bugajski [2014a] – 2 years of research – 18-element sequences, Bugajski [2014b] – 5 years

Table 2. Basic parameters of descriptive statistics characterizing the composition of treated wastewater in the analysed treatment plants

Pollution index	Parameter of the descriptive statistics	The values of the descriptive statistics for a given treatment plant		
		treatment plant A	treatment plant B	treatment plant C
BOD_5 ($mg \cdot dm^{-3}$)	maximum	27.0	45.0	42.0
	mean	12.6	14.1	15.4
	minimum	5.0	4.9	4.6
	median	12.0	12.4	14.6
	standard deviation	5.3	7.2	7.4
COD_{cr} ($mg \cdot dm^{-3}$)	maximum	123.0	116.0	121.0
	mean	58.6	62.5	69.7
	minimum	15.0	18.4	39.9
	median	56.7	62.1	67.4
	standard deviation	18.1	20.1	16.7
Total suspended solid ($mg \cdot dm^{-3}$)	maximum	34.0	35.0	33.4
	mean	16.5	18.2	18.1
	minimum	5.2	7.2	4.0
	median	15.9	17.1	18.0
	standard deviation	7.1	6.8	6.7

Explanations: BOD_5 = biochemical oxygen demands, COD_{cr} = chemical oxygen demand.
Source: own study.

research – 60-element sequences, Bugajski *et al.* [2015] – 4 years of research – 73-element sequences, Wąsik *et al.* [2016] – 3 years of research – 36-element sequences, Bugajski and Nowobilska-Majewska [2019] – 2 years of research – 87-element sequences, Kurek *et al.* [2020] – 2 years of research – 50-element sequences. As the literature review shows, the theoretical distributions were most often matched to measurement sequences with a number of 18 to 87 elements, and an average of 50 elements.

The *p*-value values presented in Tables 3–5 indicate that in the case of many analysed measurement sequences there is no reason to reject null hypotheses at the significance level of $a = 0.05$. This mainly concerns the Kolmogorov–Smirnov compliance test. For example, for a 20-element sequence of BOD_5 values, there is no reason to reject the hypotheses that each of the 9 analysed theoretical distributions correctly describes the group of these data. Only a comparison of the *p*-value of the

Table 3. The *p*-values of the probability of not rejecting the null hypothesis about the fit of a given distribution using the Kolmogorov–Smirnov concordance test (K–S) and the X^2 compatibility test for a biochemical oxygen demand

Statistical distribution	Probability (*p*-values) for different number of samples for two concordance tests							
	20		40		60		80	
	K–S	X^2	K–S	X^2	K–S	X^2	K–S	X^2
Treatment plant A								
Chi-square	0.505	**0.274**	**0.903**	**0.131**	0.578	**0.322**	0.485	0.038
Erlang	0.333	0.117	0.349	0.009	0.587	0.171	0.063	0.009
Fisher–Tippett	0.499	0.116	0.856	0.032	0.552	0.145	0.614	0.024
Gamma	0.504	0.143	0.815	0.039	0.539	0.193	0.615	**0.042**
GEV	0.378	0.039	0.291	0.005	0.488	0.101	0.617	0.012
Log-normal	0.465	0.117	0.843	0.027	0.601	0.184	**0.709**	0.040
Logistic	**0.544**	0.058	0.822	0.008	**0.619**	0.017	0.352	0.001
Normal	0.489	0.092	0.791	0.019	0.480	0.035	0.208	0.002
Weibull	0.523	0.135	0.817	0.042	0.614	0.145	0.405	0.027

cont. Tab. 3

Statistical distribution	Probability (p-values) for different number of samples for two concordance tests							
	20		40		60		80	
	K–S	X^2	K–S	X^2	K–S	X^2	K–S	X^2
Treatment plant B								
Chi-square	0.505	**0.274**	0.715	0.467	0.281	0.009	0.098	0.004
Erlang	0.333	0.117	0.617	0.154	0.271	**0.024**	0.011	0.006
Fisher–Tippett	0.499	0.116	0.463	0.231	0.295	0.006	0.252	0.139
Gamma	0.504	0.143	0.498	0.198	0.361	0.010	0.318	0.198
GEV	0.378	0.039	0.457	<0.0001	0.319	<0.0001	0.139	<0.0001
Log-normal	0.465	0.117	0.474	<0.0001	0.332	0.012	0.306	**0.316**
Logistic	**0.544**	0.058	0.682	<0.0001	0.554	0.000	0.441	0.003
Normal	0.489	0.092	0.612	<0.0001	0.505	0.000	0.388	0.001
Weibull	0.523	0.135	**0.743**	**0.743**	**0.633**	0.008	**0.588**	0.082
Treatment plant C								
Chi-square	0.957	**0.224**	0.287	0.454	0.367	<0.0001	0.424	<0.0001
Erlang	0.572	0.089	0.848	0.467	0.001	<0.0001	0.086	<0.0001
Fisher–Tippett	**0.974**	0.162	**0.931**	0.597	0.593	<0.0001	**0.782**	<0.0001
Gamma	0.956	0.209	0.603	0.553	0.260	<0.0001	0.457	<0.0001
GEV	0.672	0.051	0.922	0.409	**0.658**	<0.0001	0.723	<0.0001
Log-normal	0.953	0.192	0.733	**0.600**	0.537	<0.0001	0.753	0.000
Logistic	0.941	0.158	0.636	0.229	0.486	<0.0001	0.674	0.000
Normal	0.940	0.192	0.391	0.331	0.035	<0.0001	0.046	<0.0001
Weibull	0.912	0.164	0.313	0.308	0.079	<0.0001	0.118	<0.0001

Explanations: GEV = generalised extreme value, the highest p-value is bolded in the table.
Source: own study.

Table 4. The p-values of the probability of not rejecting the null hypothesis about the fit of a given distribution using the Kolmogorov–Smirnov concordance test (K–S) and the X^2 concordance test for a chemical oxygen

Statistical distributions	Probability (p-values) for different number of samples for two concordance tests							
	20		40		60		80	
	K–S	X^2	K–S	X^2	K–S	X^2	K–S	X^2
Treatment plant A								
Chi-square	0.966	0.016	0.076	<0.0001	0.076	<0.0001	0.003	<0.0001
Erlang	0.968	0.206	0.386	0.034	0.372	0.286	0.092	0.051
Fisher–Tippett	0.956	0.208	**0.628**	0.046	0.609	0.227	0.613	0.228
Gamma	0.932	**0.225**	0.608	0.047	**0.794**	**0.320**	0.665	0.240
GEV	0.819	0.077	0.575	0.030	0.611	0.118	0.249	0.078
Log-normal	0.972	0.216	0.663	**0.049**	0.670	0.275	0.504	**0.250**
Logistic	**0.987**	0.156	0.454	0.010	0.651	0.115	0.765	0.047
Normal	0.755	0.192	0.328	0.021	0.504	0.241	0.721	<0.0001
Weibull	0.651	0.181	0.299	0.022	0.447	0.314	**0.866**	<0.0001

cont Tab. 4

Statistical distributions	Probability (p-values) for different number of samples for two concordance tests							
	20		40		60		80	
	K–S	X^2	K–S	X^2	K–S	X^2	K–S	X^2
Treatment plant B								
Chi-square	0.966	0.016	0.021	<0.0001	0.018	<0.0001	0.008	<0.0001
Erlang	0.968	0.206	0.445	0.126	0.499	0.078	0.058	0.007
Fisher–Tippett	0.956	0.208	**0.615**	0.084	0.927	0.206	0.752	**0.290**
Gamma	0.932	**0.225**	0.494	0.136	**0.994**	**0.209**	0.931	0.222
GEV	0.819	0.077	0.229	0.004	0.954	0.153	0.269	0.111
Log-normal	0.972	0.216	0.502	0.094	0.950	0.209	0.667	0.234
Logistic	**0.987**	0.156	0.434	0.108	0.880	0.051	0.957	0.025
Normal	0.755	0.192	0.531	0.167	0.685	0.064	0.928	0.022
Weibull	0.651	0.181	0.558	**0.181**	0.833	0.127	**0.959**	0.044
Treatment plant C								
Chi-square	0.391	**0.074**	**0.976**	**0.790**	0.384	<0.0001	0.319	<0.0001
Erlang	0.515	0.037	0.640	0.440	0.200	0.123	0.855	0.233
Fisher–Tippett	**0.717**	0.047	0.718	0.688	0.724	**0.361**	0.906	**0.330**
Gamma	0.429	0.036	0.975	0.629	**0.955**	0.246	0.788	0.247
GEV	0.271	0.012	0.319	0.000	0.799	0.214	0.956	0.230
Log-normal	0.477	0.038	0.974	0.619	0.947	0.303	**0.958**	0.307
Logistic	0.363	0.015	0.830	0.332	0.852	0.090	0.772	0.085
Normal	0.334	0.026	0.846	0.519	0.667	0.062	0.353	0.029
Weibull	0.278	0.014	0.677	0.356	0.676	0.055	0.283	0.028

Explanations as in Tab. 3.
Source: own study.

Table 5. The p-values of the probability of not rejecting the null hypothesis about the fit of a given distribution using the Kolmogorov–Smirnov concordance test (K–S) and the X^2 concordance test for a total suspended solids

Statistical distribution	Probability (p-values) for different number of samples for two concordance tests							
	20		40		60		80	
	K–S	X^2	K–S	X^2	K–S	X^2	K–S	X^2
Treatment plant A								
Chi-square	0.473	0.224	0.771	0.229	0.642	0.032	0.047	0.000
Erlang	0.703	0.149	0.912	0.247	0.827	0.080	0.130	**0.130**
Fisher-Tippett	**0.940**	**0.263**	**0.974**	**0.296**	**0.981**	0.139	0.509	0.045
Gamma	0.886	0.241	0.935	0.278	0.952	**0.157**	0.534	0.082
GEV	0.893	0.154	0.705	<0.0001	0.840	0.156	**0.817**	0.020
Log-normal	0.888	0.247	0.950	<0.0001	0.943	0.144	0.558	0.042
Logistic	0.813	0.131	0.685	<0.0001	0.611	0.017	0.379	0.008
Normal	0.863	0.193	0.535	<0.0001	0.396	0.036	0.400	0.021
Weibull	–	–	–	–	–	–	–	–

cont Tab. 5

Statistical distribution	Probability (p-values) for different number of samples for two concordance tests							
	20		40		60		80	
	K–S	X^2	K–S	X^2	K–S	X^2	K–S	X^2
Treatment plant B								
Chi-square	0.473	0.224	0.716	**0.504**	0.832	**0.549**	0.205	0.032
Erlang	0.703	0.149	0.342	0.045	0.562	0.343	**0.722**	0.122
Fisher-Tippett	**0.940**	**0.263**	0.894	0.232	**0.881**	0.490	0.484	0.147
Gamma	0.886	0.241	0.848	0.236	0.714	0.377	0.359	0.119
GEV	0.893	0.154	0.136	0.038	0.740	0.365	0.442	0.108
Log-normal	0.888	0.247	**0.947**	0.221	0.849	0.481	0.555	**0.192**
Logistic	0.813	0.156	0.560	0.085	0.604	0.078	0.299	0.004
Normal	0.863	0.193	0.562	0.156	0.492	0.073	0.170	0.005
Weibull	0.671	0.161	0.570	0.169	0.714	0.114	0.313	0.028
Treatment plant C								
Chi-square	0.527	0.466	0.446	0.237	0.221	0.132	0.281	0.289
Erlang	0.640	0.599	0.216	0.098	0.118	0.104	0.100	0.325
Fisher-Tippett	0.724	0.517	0.331	0.069	0.226	0.161	0.649	0.654
Gamma	0.854	**0.662**	0.616	0.174	0.268	0.230	0.700	0.789
GEV	0.532	0.128	0.152	0.001	0.052	<0.0001	0.147	0.022
Log-normal	0.767	0.641	0.441	0.105	0.127	0.093	0.419	0.403
Logistic	0.963	0.570	0.758	0.246	**0.959**	0.194	0.586	0.691
Normal	**0.971**	0.612	0.890	0.304	0.853	0.338	0.653	0.871
Weibull	0.886	0.498	**0.939**	**0.378**	0.845	**0.404**	**0.807**	**0.947**

Explanations as in Tab. 3.
Source: own study.

rejection of individual hypotheses may indicate which distribution can be assumed with greater certainty and the smallest 1st degree error. However, it should be remembered that according to the theory of statistics, the assumption of the H_0 hypothesis is not related to the accuracy of the distribution fit.

Definitely different results regarding the fit of the theoretical and empirical distributions were obtained using the χ^2 consistency test. The results of this test narrow down the number of theoretical distributions that reflect empirical distributions. However, with measurement sequences of up to 40 elements, there is still no reason to reject several distributions of different shapes. To avoid the error of selecting the wrong hypothesis H_0, the probability values of p-value should be taken into account.

As Tables 3–5 contain a very large amount of data (324 statistical analyses), Table 6 summarises the total number of times when selecting a given theoretical distribution to empirical data, the H_0 hypothesis at the significance level of $a = 0.05$ was not rejected. If a given theoretical distribution reflects the data distribution of 1 pollution index for each of the 4 analysed communities and additionally 3 treatment plants, then the total value given in Table 6 is 12. If a given theoretical distribution describes all the tested pollution indicators, then the sum given in Table 6 is 36.

The results presented in Table 6 undermine the reliability of the K–S conformance test for the selection of the distribution. On the basis of this test, for BOD_5 – 7 out of 9 analysed distributions correctly describe the empirical data, for COD_{Cr} – 8 out of 9, and for TSS – 9 out of 9. It is difficult to accept such results, taking into account the different shape of the analysed theoretical distributions. Based on this test, one can only point out how high the probability there is no basis for rejecting the H_0 hypothesis.

Using the χ^2 compatibility test, it was found that the BOD_5 values in treated sewage most often describe the Chi-square, Fisher–Tippett, Gamma and Weibull distributions. The probability of not rejecting the Chi-square distribution (Tab. 3) was the highest in 5 out of 12 analyses. COD_{Cr} values were most often described by the Fisher–Tippett distribution (12 times out of 12 analyses). However, in the case of this indicator (Tab. 4), the highest probabilities of not rejecting the H_0 hypothesis were obtained for the Gamma distribution (in 4 out of 12 analyses). The concentrations of TSS were most often described by the following distributions: Erlang, Fisher–Tippett and Gamma. The highest probabilities of not rejecting the H_0 hypothesis were obtained for the Fisher–Tippett distributions (in 3 out of 12 analyses) and Weibull (also in 3 out of 12 analyses).

Table 6. Summary indicating the number of cases in which a given theoretical distribution describes the empirical distribution of the value of a given indicator of pollution with a probability equal to or greater than 95%

Statistical distribution	Number of cases for investigated pollution index for two concordance tests							
	BOD_5		COD_{Cr}		total suspended solid		total	
	K–S	X^2	K–S	X^2	K–S	X^2	K–S	X^2
Chi-square	12	7	8	2	12	9	32	18
Erlang	10	6	12	9	12	12	34	27
Fisher–Tippett	12	7	12	12	12	12	36	31
Gamma	12	7	12	11	12	12	36	30
GEV	12	3	12	8	12	6	36	17
Log-normal	12	6	12	11	12	10	36	27
Logistic	12	4	12	9	12	8	36	21
Normal	11	4	12	7	12	8	35	19
Weibull	12	7	12	7	12	10	36	24

Explanations: BOD_5 = biochemical oxygen demands, COD_{cr} = chemical oxygen demand; GEV = generalised extreme value.
Source: own study.

Taking into account all the analysed values of pollution indicators, according to the χ^2 test, these communities were most often described by the Fisher–Tippett distribution (31 out of 36 analyses) and Gamma (30 out of 36 analyses). On the other hand, taking into account the probability of not rejecting the H_0 hypothesis, the greatest certainty of its acceptance was obtained for the Chi-square distribution (9 out of 36 analyses). This is valuable information because the distribution is a transformation of the normal distribution by taking the measurement sequence data to the second power. The Chi-square distribution is a non-negative, right-hand asymmetric distribution. Chi-square data can be easily transformed to a normal distribution if necessary or useful for in-depth statistical analysis using parametric tests.

The analysis of the matching of distributions, depending on the number of data in the measurement series (Tabs. 3–5) showed that the highest values of non-rejection of the H_0 hypothesis in the χ^2 test were obtained for measurement sequences with a 40-element number for COD_{Cr} and total suspension, and for an 80-element number for BOD_5.

Młyński *et al.* [2019] using the function, made theoretical distributions fitting the distributions of empirical pollution indicators: distribution of GEV, GMM, log-normal, normal, Pareto, Rayleigh, triangular and Weibull. They assessed the compatibility of distributions using the Anderson–Darling (A–D) test for the significance level $\alpha = 0.05$. The authors found that the values of BOD_5 and COD_{Cr} in treated wastewater were best described by GMM decomposition, while *TSS* by GEV decomposition. The results obtained in the work of Młyński *et al.* [2019] differ from the results in this paper, but it may result from the difference in the size of the research object. In the case of large wastewater treatment plants (with more than 2000 PE), the requirements of the water permit regarding the quality of treated wastewater increase. Hence, empirical data are grouped into intervals with a smaller range. It certainly influenced the shape of the empirical and theoretical statistical distribution of these data.

The objects most similar, in terms of total population equivalent (PE), were analysed by Bugajski *et al.* [2012; 2016], Nastawny and Jucherski [2013], Bugajski [2014a, b], and Marzec [2017]. The cited studies found that the values of BOD_5, COD_{Cr} and total suspension were best described by the Weibull distribution. Unfortunately, these studies did not provide the *p*-value probability of not rejecting the H_0 hypothesis. It was also not mentioned which significance test was used to assess the fit of the theoretical to the empirical distribution.

In other studies (Niku *et al.* [1981], Andraka [2005; 2011; 2020], and Oliveira *et al.* [2012]) concerning data from much larger sewage treatment plants than 2000 PE, the authors concluded that the values of pollutants in treated sewage were best described by the log-normal distribution. In these works, it is also difficult to assess the level of matching of theoretical distributions to empirical data, because the authors did not provide the results of significance tests. There was also no information as to whether other theoretical distributions were tested.

On the basis of the presented discussion of the results, it is difficult to indicate one theoretical distribution, most often describing the values of BOD_5, COD_{Cr} and total suspended solids in treated sewage. However, according to the conducted own research and the obtained results, it was established that this distribution is Fisher–Tippett. On the basis of the χ^2 test, this distribution was described by a total of 31 out of 36 analysed groups at the significance level of $a = 0.05$. Other distributions that often describe the analysed empirical data are: Gamma, log-normal, Chi-square, and Weibull. The common feature of these distributions is usually asymmetry, right oblique. The skewness value calculated on the basis of the data in Table 2 ranges from 0.15 to 1.69.

CONCLUSIONS

1. The theoretical distribution most often describing the empirical distribution of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD_{Cr}) values and total suspended solids in treated sewage, discharged from small wastewater treatment plants with PE < 2000 – is the Fisher–Tippett distribution.
2. A greater certainty of the correctness of the theoretical distribution to the empirical data was obtained using the χ^2 significance test than with the Kolmogorov–Smirnov (K–S) test.
3. The analysis of the theoretical distribution fit, depending on the number of data in the measurement series, showed that higher probabilities of not rejecting the H_0 hypothesis in the χ^2 test were obtained for measurement sequences with a 40-element number for COD_{Cr} and total suspended solids, and for an 80-element number for BOD_5.
4. The values of BOD_5, COD_{Cr} and total suspended solids in treated sewage are most often described by an asymmetric,

right-oblique distribution. This precludes the direct use of parametric tests in statistical analyses.
5. The research showed that the K–S test does not provide sufficient certainty as to the correctness of the theoretical statistical distribution to the empirical data.
6. To obtain appropriate reliability of the goodness of fit of the theoretical distribution to the empirical distribution, as many theoretical distributions as possible should be tested using different significance tests.

REFERENCES

Andraka D. 2005. Wykorzystanie statystycznej kontroli jakości do oceny pracy oczyszczalni ścieków [The use of statistical quality control to evaluate the operation of sewage treatment plants]. Monografie Komitetu Inżynierii Środowiska PAN. Nr 30 p. 565–580.

Andraka D. 2011. Modelowanie pracy oczyszczalni ścieków z wykorzystaniem symulacji Monte Carlo [Modeling of wastewater treatment plant operation by means of Monte Carlo simulation]. Inżynieria Ekologiczna. Nr 24 p. 7–16.

Andraka D. 2020. Reliability analysis of activated sludge process by means of biokinetic modelling and simulation results. Water. Vol. 12(1), 291. DOI 10.3390/w12010291.

Andraka D., Dzienis L. 2003. Wymagany poziom niezawodności oczyszczalni ścieków w świetle przepisów polskich i europejskich [Required reliability level of wastewater treatment plants according to European and Polish regulations]. Zeszyty Naukowe Politechniki Białostockiej. Inżynieria Środowiska. Z. 16 p. 24–28.

Bugajski P. 2014a. Analysis of reliability of the treatment plant Bioblok PS-50 using the method of Weibull. Infrastruktura i Ekologia Terenów Wiejskich. Nr 3 p. 667–677. DOI 10.14597/infraeco.2014.2.2.049

Bugajski P. 2014b. Ocena niezawodności usuwania związków biogennych w oczyszczalni ścieków metodą Weibulla [Assessment of nutrient removal reliability in a sewage treatment plant using the Weibull method]. Zeszyty Problemowe Postępów Nauk Rolniczych. Z. 576 p. 13–21. [Access 15.01.2022]. Available at: http://zppnr.sggw.pl/576-02.pdf

Bugajski P., Almeida Araujo M.A., Kurek K. 2016. Reliability of sewage treatment plants processing sewage from school buildings located in non-urban areas. Infrastruktura i Ekologia Terenów Wiejskich. Nr VI/3 p. 1547–1557. DOI 10.14597/infraeco.2016.4.3.115.

Bugajski P., Kaczor G., Bergel T. 2015. Niezawodność usuwania azotu ze ścieków w zbiorczej oczyszczalni z sekwencyjnym reaktorem biologicznym [The removal of reliability nitrogen in wastewater treatment plant witch sequencing biological reactor]. Acta Scientiarum Polonorum. Formatio Circumiectus. Vol. 14(3) p. 19–27. DOI 10.15576/ASP.FC/2015.14.3.19.

Bugajski P., Nowobilska-Majewska E. 2019. A Weibull analysis of the reliability of a wastewater treatment plant in Nowy Targ, Poland [online]. Rocznik Ochrona Środowiska. Vol. 21 p. 825–840. Available at: https://ros.edu.pl/images/roczniki/2019/050_ROS_V21_R2019.pdf

Bugajski P., Pawełek J., Kurek K. 2017. Concentrations of organic and biogenic pollutants in domestic wastewater after mechanical treatment in the aspect of biological reactor design. Infrastruktura i Ekologia Terenów Wiejskich. Nr IV/3 p. 1811–1822. DOI 10.14597/infraeco.2017.4.3.136.

Bugajski P., Wałęga A., Kaczor G. 2012. Zastosowanie metody Weibulla do analizy niezawodności działania przydomowej oczyszczalni ścieków [Application of Weibull method to analyze reliability of operation of a household sewage treatment plant]. Gaz, Woda i Technika Sanitarna. Nr 2 p. 56–58.

Chmielowski K., Bugajski P., Wąsik E. 2015. Ocena działania oczyszczalni ścieków w Haczowie przed i po modernizacji [Assessment of the operation of a sewage treatment plant in Haczów before and after modernization]. Infrastruktura i Ekologia Terenów Wiejskich. Nr IV/1 p. 949–964. DOI 10.14597/infraeco.2015.4.1.076.

Chmielowski K., Satora S., Wałęga A. 2009. Ocena niezawodności działania oczyszczalni ścieków dla gminy Tuchów [Evaluation of the reliability of the sewage treatment plant for the commune of Tuchów]. Infrastruktura i Ekologia Terenów Wiejskich. Nr 9 p. 63–72.

Kurek K., Bugajski P., Operacz A., Śliz P., Jóźwiakowski K., Almeida A. 2020. Reliability assessment of pollution removal of wastewater treatment plant using the method of Weibull. In: The 9th International Scientific-Technical Conference on Environmental Engineering, Photogrammetry, Geoinformatics – Modern Technologies and Development Perspectives (EEPG Tech 2019). Vol. 171, 01007 DOI 10.1051/e3sconf/202017101007.

Luszniewicz A., Słaby T. 2003. Statystyka z pakietem komputerowym STATISTICA PL [Statistics with the STATISTICA computer package]. Warszawa. Wydawnictwo C.H. Beck. ISBN 83-7247-798-1 pp. 445.

Marzec M. 2017. Reliability of removal of selected pollutants in different technological solutions of household wastewater treatment plants. Journal of Water and Land Development. No. 35 p. 141–148. DOI 10.1515/jwld-2017-0078.

Meijer D., Van Bijnen M., Langeveld J., Korving H., Post J., Clemens F. 2018. Sewer networks using graph-theory. Water. Vol. 10, 136. DOI 10.3390/w10020136.

Młyńska A., Chmielowski K., Młyński D. 2017. Analiza zmian jakości ścieków w trakcie procesów oczyszczania na oczyszczalni w Przemyślu [The analysis of the changes in the sewage quality during treatment processes on the wastewater treatment plant in Przemyśl]. Inżynieria Ekologiczna. Vol. 18(5) p. 18–26. DOI 10.12912/23920629/74973.

Młyński A., Chmielowski K., Młyński D. 2016a. Ocena skuteczności oraz stabilności pracy oczyszczalni ścieków w Zabajce [The assesment of the efficency and stability of work sewage treatment plant in Zabajka]. Inżynieria Ekologiczna. Vol. 47 p. 123–130. DOI 10.12912/23920629/62856.

Młyński D., Bugajski P., Młyńska A. 2019. Application of the mathematical simulation methods for the assessment of the wastewater treatment plant operation work reliability. Water. Vol. 11, 873. DOI 10.3390/w11050873.

Młyński D., Chmielowski K., Młyńska A., Miernik W. 2016b. Ocena skuteczności pracy oczyszczalni ścieków w Jaśle [Evaulation of efficency of sewage treatment plant in Jasło]. Infrastruktura i Ekologia Terenów Wiejskich. Vol. I/1 p. 147–162. DOI 10.14597/infraeco.2016.1.1.011.

Młyński D., Młyńska A., Chmielowski K., Pawełek J. 2020. Investigation of the wastewater treatment plant processes efficiency using statistical tools. Sustainability. Vol. 12(24), 10522. DOI 10.3390/su122410522.

Nastawny M., Jucherski A. 2013. Ocena technologicznej niezawodności przydomowej oczyszczalni ścieków z układem złóż filtracyjnych zmodyfikowaną metodą Weibulla [Assessing technical reliability of an on-site sewage treatment plant with filtration bed system, by using modified Weibull's method]. Problemy Inżynierii Rolniczej. Nr 2(80) p. 165–175.

Niku S., Schroeder E.D., Tchobanoglous G., Samaniego F.J. 1981. Project summary. In: Performance of activated sludge processes: Reliability, stability and variability. EPA-600/S2-81-227. Cincinnati, OH, USA. EPA. Vol. 53(5) p. 546–559.

Oliveira S.C., Souki I., von Sperling N. 2012. Lognormal behaviour of untreated and treated wastewater constituents. Water Science & Technology. Vol. 63 p. 596–603. DOI 10.2166/wst.2012.899.

Oliveira S.C., von Sperling M. 2008. Reliability analysis of wastewater treatment plants. Water Research. Vol. 42(4–5) p. 1182–1194. DOI 10.1016/j.watres.2007.09.001.

PN-EN 872:2007 – wersja polska. Jakość wody – oznaczanie zawiesin – Metoda z zastosowaniem filtracji przez sączki z włókna szklanego [Polish version. Water quality – determination of suspended solids – Method using filtration through glass fiber filters]. Warszawa. PKN pp. 12.

PN-EN ISO 5815-1:2019-12. Jakość wody – oznaczanie biochemicznego zapotrzebowania tlenu po n dniach (BZTn) – Część 1: Metoda rozcieńczeń, z dodatkiem materiału zaszczepiającego i allilotiomocznika [Water quality – Determination of biochemical oxygen demand after n days (BODn) – Part 1: Dilution method, with addition of seed and allylthiourea]. Warszawa. PKN pp. 36.

PN-ISO 6060: 2006. Jakość wody – oznaczanie chemicznego zapotrzebowania tlenu [Water quality – Determination of chemical oxygen demand]. Warszawa. PKN pp. 10.

Rozporządzenie Ministra Gospodarki Morskiej i Żeglugi Śródlądowej z dnia 12 lipca 2019 r. w sprawie substancji szczególnie szkodliwych dla środowiska wodnego oraz warunków, jakie należy spełnić przy wprowadzaniu do wód lub do ziemi ścieków, a także przy odprowadzaniu wód opadowych lub roztopowych do wód lub do urządzeń wodnych [The Regulation of the Minister of Maritime Economy and Inland Navigation of 12 July 2019 on the substances that are particularly harmful to the aquatic environment and the conditions to be met upon discharging them into water or ground, and upon discharging rain water and thaw water onto the water or to water facilities]. Dz.U. 2019 poz. 1311.

Siwiec T., Reczek L., Michel M.M, Gut B, Hawer-Strojek P., Czajkowska J., Jóźwiakowski K., Gajewska M., Bugajski P. 2018. Correlations between organic pollution indicators in municipal wastewater. Archives of Environmental Protection. Vol. 44(4) p. 50–57. DOI 10.24425/aep.2018.122296.

Śliz P. 2018. Analiza skuteczności, niezawodności i stabilności procesów w Oczyszczalni Ścieków Kraków-Płaszów [An analysis of effectiveness, reliability and process stability at Krakow-Płaszow Sewage Treatment Plant]. Świat Nieruchomości. T. 105(3) p. 77–82.

Sobczyk M. 2022. Statystyka [Statistics]. Warszawa. Wydaw. Nauk. PWN. ISBN 9788301151997 pp. 428.

Taheriyoun M., Moradinejad S. 2015. Reliability analysis of a wastewater treatment plant using fault tree analysis and Monte Carlo simulation. Environmental Monitoring and Assessment. Vol. 187 p. 4186–4199. DOI 10.1007/s10661-014-4186-7.

Tchórzewska-Cieślak B., Piegdoń I. 2016. The method of identification the failure risk on water supply networks. Journal of KONBiN. Vol. 1(37) p. 73–94. DOI 10.1515/jok-2016-0004.

Van Riel W., van Bueren E., Langeveld J., Herder P., Clemens F. 2015. Decision-making for sewer asset management: Theory and practice. Urban Water Journal. Vol. 13 p. 57–68. DOI 10.2166/wst.2016.253

Wałęga A. 2009. Ocena funkcjonowania oczyszczalni ścieków metodami statystycznymi [Assessment of the functioning of wastewater treatment plants using statistical methods]. Forum Eksploatatora. Nr 5(44) p. 30–34.

Wąsik E., Bugajski P., Chmielowski K. 2016. Model Weibulla jako narzędzie oceny niezawodności działania oczyszczalni ścieków w Niepołomicach [Weibull model as a tool for assessment of operation reliability in a sewage treatment plant in Niepołomice]. Nauka, Przyroda, Technologie. Nr 10(2) #20 p. 1–11. DOI 10.17306/J.NPT.2016.2.20.

Zawadzka B., Siwiec T., Marzec M. 2021. Effectiveness of dairy and domestic wastewater treatment and technological reliability of the wastewater treatment plant in Michów. Poland. Journal of Ecological Engineering. Vol. 22(10) p. 141–151. DOI 10.12911/22998993/1421.