ZESZYTY NAUKOWE POLITECHNIKI RZESZOWSKIEJ Nr 276
Budownictwo i Inżynieria Środowiska z. 58 (4/11) 2011

Dariusz ANDRAKA
Politechnika Białostocka

WYKORZYSTANIE NARZĘDZI STATYSTYCZNYCH W PROCESIE PROJEKTOWANIA OCZYSZCZALNI ŚCIEKÓW

Procesowi projektowania oczyszczalni ścieków, na jego różnych etapach, nierozłącznie towarzyszy proces podejmowania decyzji – począwszy od przyjęcia miarodajnych danych wyjściowych, kończąc na wyborze optymalnego wariantu rozwiązań techniczno-technologicznych. W niniejszym artykule zaprezentowano różne metody statystycznej analizy i oceny zgromadzonych danych wyjściowych charakteryzujących ścieki dopływające do oczyszczalni. Celem analizy jest ustalenie miarodajnych parametrów projektowych z jak najmniejszym błędem oszacowania. Rozważania teoretyczne zostały poparte przykładami praktycznego zastosowania przedstawionych w referacie metod i technik statystycznych. Do weryfikacji uzyskanych wyników wykorzystano model symulacyjny dopływu ścieków do oczyszczalni, stosujący metodę Monte-Carlo. Prezentowana praca jest wynikiem badań prowadzonych przez autora w ramach pracy statutowej S/WBiIŚ/22/08 realizowanej w Katedrze Systemów Inżynierii Środowiska Politechniki Białostockiej.

1. Wprowadzenie

W procesie projektowania komunalnych oczyszczalni ścieków wykorzystujących osad czynny projektant musi podejmować decyzje związane z wyborem optymalnego rozwiązania zadania projektowego – począwszy od danych wyjściowych przyjmowanych do projektowania, poprzez adekwatność zastosowanych modeli obliczeniowych, a kończąc na ocenie uzyskanych wyników. Oprócz czynników „niewymiernych" (doświadczenie, intuicja) oraz tradycyjnych (wiedza, analogia do rozwiązań już istniejących) proces podejmowania decyzji może być wspomagany narzędziami bardziej wymiernymi, opartymi przede wszystkim na analizie statystycznej dostępnych danych i wyników.

Jednym z podstawowych czynników decydujących o poprawności zastosowanych rozwiązań może być przyjęcie odpowiednio dobranego zestawu danych charakteryzujących ścieki dopływające do oczyszczalni, będącego następnie podstawą wymiarowania obiektów oczyszczalni. Mając do dyspozycji nawet niezbyt liczny zbiór danych pochodzących z badań własnych prowadzonych na oczyszczalni czy też monitoringu WIOŚ oraz stosując odpowiednie metody statystyczne, można określić miarodajne dane projektowe. Pozwalają one na właściwy dobór parametrów techniczno-technologicznych projektowanych obiektów, co zapewnia spełnienie wymagań przepisów określających warunki, jakim muszą odpowiadać ścieki oczyszczone.

2. Metody analizy dopływów do oczyszczalni

W Polsce najczęściej stosowaną metodą obliczeniową jest procedura opisana w arkuszu roboczym ATV-A131 [1], w którym wykorzystano statyczne modele procesów nitryfikacji, denitryfikacji i rozkładu węgla organicznego, realizowane dla określonego stanu przyjętego do obliczeń. W metodzie tej nie uwzględnia się w sposób bezpośredni wahań stężeń i ładunków zanieczyszczeń w dopływie, natomiast bierze się pod uwagę różne warianty obciążenia oczyszczalni, stosując odpowiednie współczynniki bezpieczeństwa. Dlatego też dla przyszłego działania oczyszczalni szczególnie ważne jest przyjęcie miarodajnego obciążenia oczyszczalni (zarówno hydraulicznego, jak i ładunkiem zanieczyszczeń), pozwalającego na prawidłowe funkcjonowanie w różnych warunkach eksploatacyjnych. Według niemieckich materiałów źródłowych zaleca się wyznaczenie co najmniej 3-miesięcznych badań (najlepiej z uwzględnieniem przypadków najmniej korzystnych, występujących z reguły w okresie zimowo-wiosennym), arkusz ATV-A131 wymaga zaś badań z 9 miesięcy (dla określenia wartości do wymiarowania oczyszczalni), przy czym konieczne jest także uruchomienie instalacji półtechnicznej w warunkach zbliżonych do rzeczywistych [1, 2]. W polskich warunkach wymagania takie są często trudne do spełnienia, jednakże projektant powinien dołożyć wszelkich starań, aby zebrać możliwie obszerny i reprezentatywny zbiór danych wstępnych i na ich podstawie oszacować parametry projektowe. Należy przy tym pamiętać, że wielkości miarodajne do wymiarowania obiektów oczyszczalni powinny uwzględniać zmienność rzeczywistej charakterystyki dopływu do oczyszczalni, a nie tylko jego uśrednioną wartość, możliwą do określenia na podstawie wskaźników jednostkowych (które to w poprzednich latach były nadmiernie „eksploatowane” przez polskich projektantów). Wartości istotne dla projektowania różnych elementów oczyszczalni zestawiono w tab. 1.

Ogólnie można stwierdzić, że kluczową wartością dla większości parametrów projektowych jest zmienna odpowiadająca skumulowanemu prawdopodobieństwu 85% występowania w danej zbiorowości. Z punktu widzenia statystyki matematycznej parametr ten określany jest mianem 85. percentyla rozkładu zmiennej (P_{85}) i jest on argumentem funkcji dystrybuanty F_n (skumulowanego prawdopodobieństwa) rozkładu zmiennej losowej. Analizując dany rozkład empiryczny zmiennej losowej X przedziałami (tzn. wyznaczając jego dystrybuantę empiryczną), łatwo można obliczyć wartość dowolnego percentyla rzędu p (P_p):

$$Fn\left(P_p\right) \geq p, \qquad P_p = x_{0p} + \left[p - F_n\left(x_{0p}\right) \right] \frac{h_p}{w_p} \tag{1}$$

gdzie: p – rząd percentyla ($0 < p < 1$),
x_{0p} – dolna granica przedziału, w którym występuje percentyl P_p,
$F_n(x_{0p})$ – skumulowana częstość względna dla dolnej granicy przedziału, w którym znajduje się percentyl P_p (dystrybuanta empiryczna dla dolnej granicy przedziału),
h_p, w_p – odpowiednio częstość i rozpiętość przedziału percentyla P_p.

Tabela 1. Parametry miarodajne do wymiarowania oczyszczalni z osadem czynnym

Parametr	**Cel obliczeń**	**Wariant obliczeń**	**Wielkość miarodajna**
Q_{bd} – obliczeniowy dopływ ścieków	osad czynny	kanalizacja ogólnospławna	85% wartości dla dni bez deszczu
		kanalizacja rozdzielcza	99% wartości wszystkich dni
Q_{bd} max	osadniki wtórne	linia przebiegu dobowego, godzinowy dopływ szczytowy	
Ładunek BZT_5	wiek osadu	bez nitryfikacji	85% wartości wszystkich dni roboczych
		z nitryfikacją	średnia z tygodnia o największym obciążeniu (lub 85% wartości wszystkich dni)
	doprowadzenie tlenu	bez nitryfikacji	85% wartości z wszystkich dni roboczych
		z nitryfikacją	linie przebiegu dobowego (85% wartości z wszystkich dni)
	osad nadmierny	–	50% wartości (mediana)
SM_{org} – sucha masa organiczna	osad nadmierny	–	50% wartości (mediana)
TKN – azot ogólny Kiejdahla	wymiarowanie komór	z nitryfikacją i denitryfikacją	85% wartości z wszystkich dni
	doprowadzenie tlenu	z nitryfikacją	linie przebiegu dobowego (85% wartości z wszystkich dni)

Jeszcze prostszym rozwiązaniem jest zebranie dostępnych danych w arkuszu kalkulacyjnym i skorzystanie z wbudowanych w program gotowych funkcji statystycznych. W ten sposób można oszacować wartość parametrów, które stanowią górną granicę 85% przypadków (inaczej mówiąc nie zostaną przekroczone częściej niż w 15% przypadków), przy czym dokładność estymacji będzie tym większa, im większa będzie liczebność zebranych danych.

Wobec tego powstaje pytanie – na ile oszacowanie jest dokładne, jeżeli dysponuje się niewielką liczbą pomiarów i czy przyjęcie wartości miarodajnych do wymiarowania oczyszczalni nie będzie obarczone zbyt dużym błędem? Do oszacowania tego błędu mogą posłużyć wyznaczone wartości błędu standardowego, który jest funkcją odchylenia standardowego i liczby obserwacji. Istnieje kilka możliwych przypadków błędów [3]:

- błąd estymatora wartości średniej m_x (dla N > 30 lub rozkładu normalnego)

$$\sigma_{m_x} = \frac{\sigma}{\sqrt{N}} \tag{2}$$

- błąd dowolnego estymatora Θ (dla rozkładu normalnego)

$$\sigma_{\Theta} = \frac{2\sigma}{\sqrt{N}} \tag{3}$$

- błąd dowolnego estymatora Θ (dla rozkładu innego niż normalny)

$$\sigma_{\Theta} = \frac{4\sigma}{\sqrt{N}} \tag{4}$$

Aby jednak ustalić wiarygodną wartość parametrów projektowych na podstawie dostępnego zbioru danych wyjściowych, można spróbować dopasować jeden z typowych rozkładów statystycznych do rozkładu empirycznego badanej zmiennej i skorzystać ze znanych zależności do wyznaczenia niezbędnych parametrów. Szczególne znaczenie mają w tym przypadku rozkłady normalny i logarytmiczno-normalny. Rozkład normalny występuje powszechnie w przyrodzie i opisuje zmienne, których wielkość zależy od sumy (lub średniej) wielu drobnych losowych czynników. Z kolei rozkład logarytmiczno-normalny mają zmienne, których logarytm (standardowo naturalny) ma rozkład normalny. Wartość tych zmiennych jest często wynikiem multiplikatywnego działania wielu drobnych czynników losowych (w odróżnieniu od addytywnego wpływu podobnych czynników na zmienną o rozkładzie normalnym). Warto również pamiętać o tym, że – zgodnie z *centralnym twierdzeniem granicznym* – przy rosnącej liczebności próby jej rozkład statystyczny dąży do rozkładu normalnego (nawet gdy badana zmienna nie ma rozkładu normalnego). W związku z tym w wielu przypadkach założenie o normalności rozkładu zmiennej losowej (lub jej logarytmu) może być uzasadnione, gdy wstępna analiza danych (zwłaszcza w próbce o niewielkiej liczebności) nie wskazuje na taki rozkład zmiennej.

Rozkład normalny w postaci standardowej charakteryzuje się średnią $\mu_s = 0$ oraz odchyleniem standardowym $\sigma_s = 1$, co zapisuje się $N(0,1)$. Rozkład ten charakteryzuje się wieloma ciekawymi właściwościami matematycznymi, co sprawia, że metody statystyczne związane z jego zastosowaniem są dosyć proste obliczeniowo (m.in. poprzez łatwą dostępność do stablicowanych wartości funkcji dystrybuanty). Dodatkowo dla dowolnej zmiennej losowej X o rozkładzie normalnym $N(\mu,\sigma)$ istnieje zależność:

$$F(x) = P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(z) = p \tag{5}$$

gdzie: $F(x)$ – dystrybuanta zmiennej losowej X o rozkładzie $N(\mu,\sigma)$,
x – wartość zmiennej losowej X,
$\Phi(z)$ – dystrybuanta rozkładu $N(0,1)$,
z – wartość zmiennej losowej X poddana standaryzacji,

$$z = \frac{x-\mu}{\sigma} \tag{6}$$

Wprowadzając do równania (6) parametr nazywany współczynnikiem zmienności i obliczany ze wzoru:

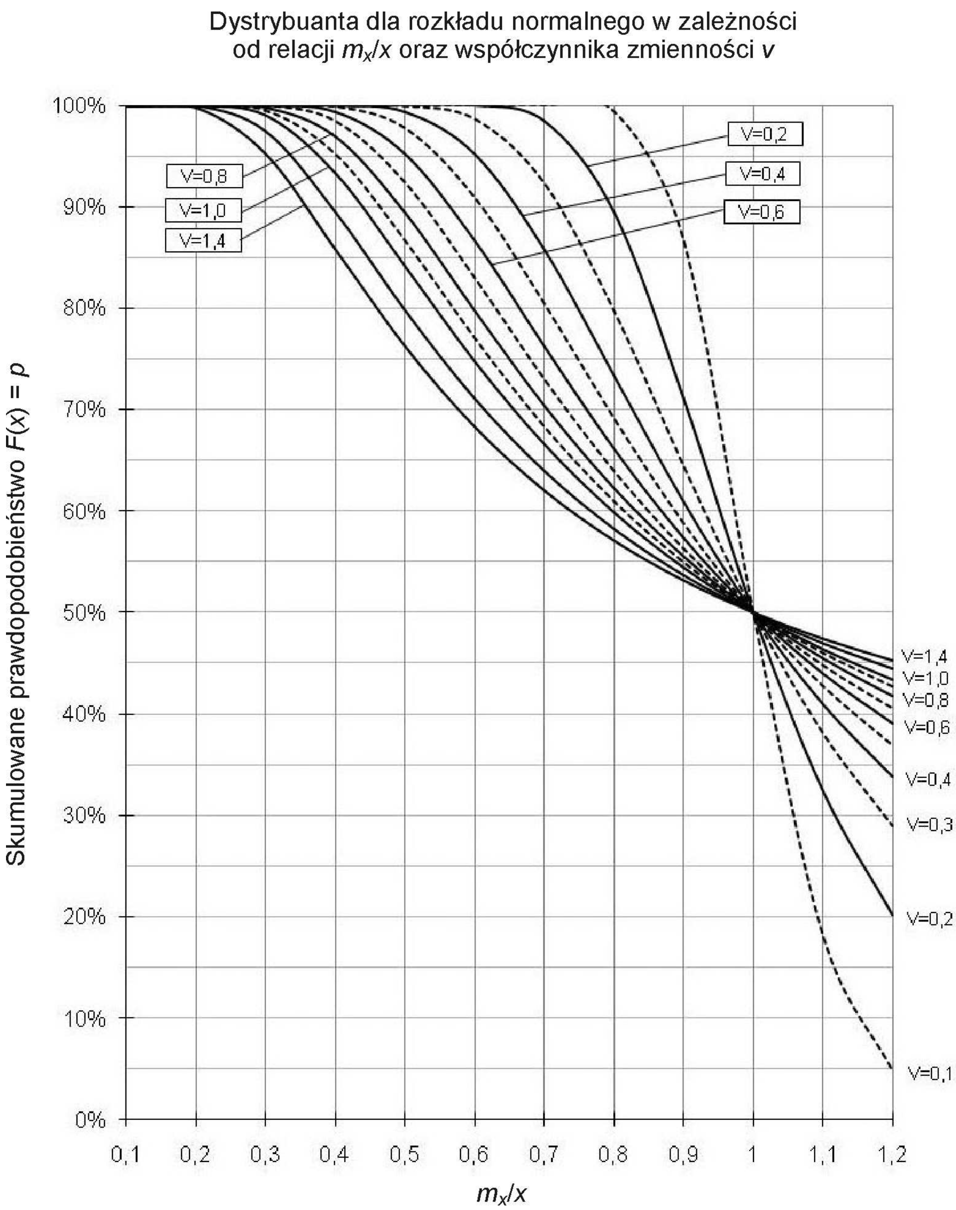
$$\nu = \frac{\sigma}{\mu} \tag{7}$$

otrzymuje się zależność pozwalającą modelować relacje pomiędzy wartościami funkcji dystrybuanty a wartościami zmiennej losowej dla rozkładów normalnych o różnych parametrach (zdeterminowanych wartością współczynnika zmienności):

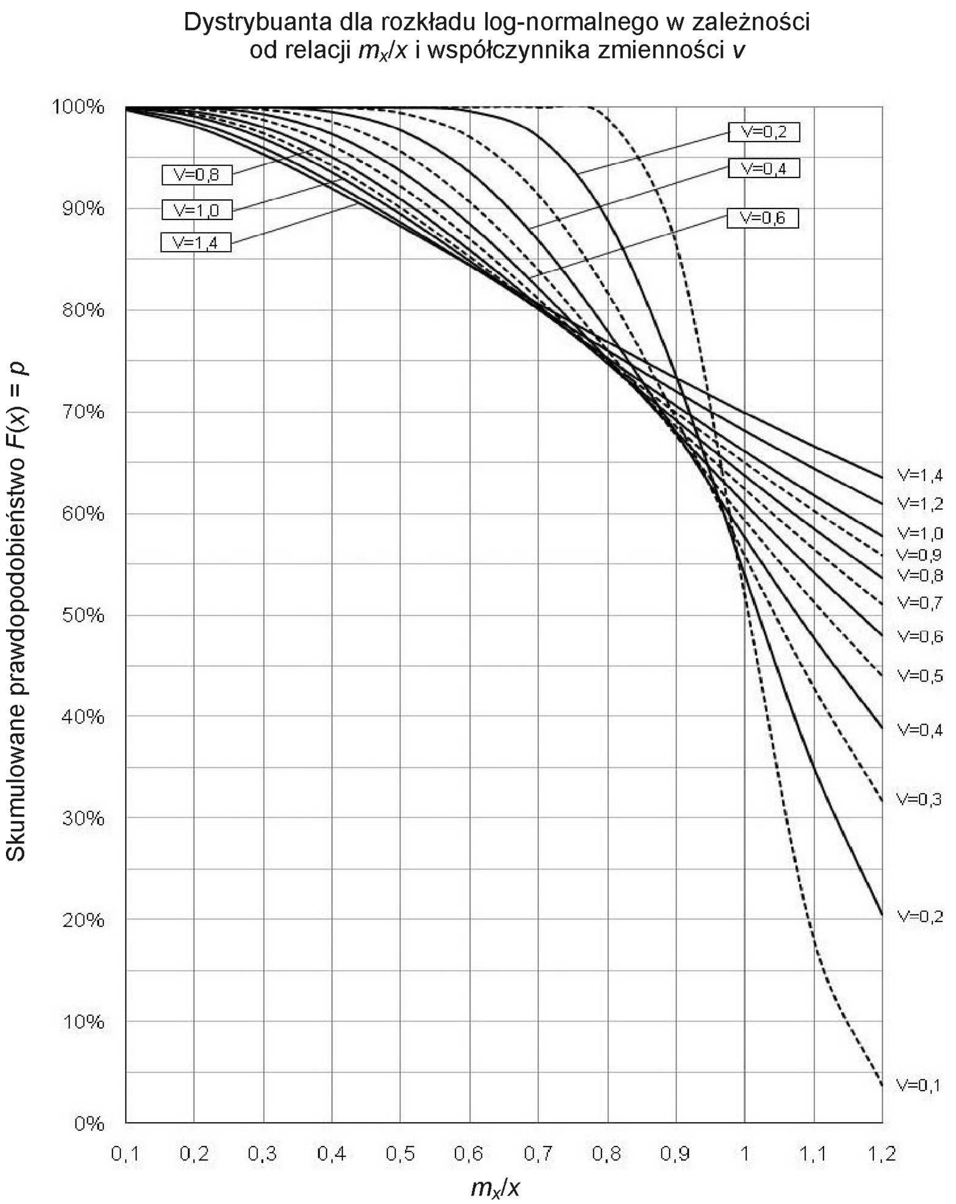
$$N(\mu,\sigma): z = \frac{1-\frac{\mu}{x}}{\frac{\mu}{x}\cdot\nu} \quad \text{i} \quad \Phi(\mathrm{z}) = p \tag{8}$$

W analogiczny sposób można wyprowadzić zależność rozkładu logarytmiczno-normalnego:

$$\ln N(\mu,\sigma): z = \ln\left(\frac{\mu}{x}\frac{1}{\sqrt{\nu^{2+1}}}\right) / \sqrt{\ln\left(\nu^2+1\right)} \quad \text{i} \quad \Phi(\mathrm{z}) = p \tag{9}$$



Rys. 1. Nomogram do wyznaczania dystrybuanty lub percentyla rozkładu normalnego przy różnych wartościach współczynnika zmienności



Rys. 2. Nomogram do wyznaczania dystrybuanty lub percentyla rozkładu log-normalnego przy różnych wartościach współczynnika zmienności

Na podstawie równań (8) oraz (9) zostały sporządzone nomogramy (rys. 1. i 2.), za pomocą których można wyznaczać m.in. wartości percentyla rzędu p (P_p) dla rozkładu normalnego (lub log-normalnego) o parametrach (μ, σ), odpowiadającego – zgodnie z równaniem (5) – wartości zmiennej losowej $X = x$, której dystrybuanta $F(x) = p$. Parametry rozkładu statystycznego zastępuje się ich estymatorami wyznaczonymi z próby rzeczywistej – m_x dla średniej μ oraz s_x dla odchylenia standardowego σ.

Sposób korzystania z nomogramów jest stosunkowo prosty. Zakłada się, że badana zmienna losowa X ma rozkład logarytmiczno-normalny i są dla niej wyznaczone statystyki opisowe o następujących wartościach: średnia $m_x = 1000$, odchylenie standardowe $s_x = 300$, współczynnik zmienności $v = 0{,}3$. Szukane są wartości percentyla P_{85} dla tej zmiennej.

Na rysunku 2. na osi rzędnych odszukuje się wartość skumulowanego prawdopodobieństwa $p = 85\%$ i prowadzi linię poziomą w prawo do punktu przecięcia z linią rozkładu o współczynniku zmienności $v = 0{,}3$. Z tego punktu należy poprowadzić linię pionową w dół i na osi odciętych odczytać wartość relacji m_x/x. Dla danych przykładowych wynosi ona ok. 0,77. Na tej podstawie można obliczyć wartość zmiennej losowej $x = m_x/0{,}77 = 1000/0{,}77 = 1299$. Odpowiada ona wartości percentyla P_{85}, co oznacza, że analizowana zmienna nie powinna przekroczyć wartości 1299 w 85% przypadków.

3. Praktyczne aspekty wyboru parametrów projektowych

Do oceny przydatności zaprezentowanych w poprzednim punkcie narzędzi statystycznych przeanalizowano 3 zbiory danych, pochodzące z obiektów o różnej wielkości i odmiennej specyfice systemów kanalizacyjnych. Oczyszczalnia *A* obsługuje miasto o wielkości ok. 300 tys. mieszkańców i gromadzi ścieki komunalne z niewielkim udziałem ścieków przemysłowych. Oczyszczalnia *B* odbiera ścieki od ok. 35 tys. mieszkańców, z dużym udziałem ścieków przemysłowych. Oczyszczalnia *C* obejmuje obszar funkcjonalny zamieszkany przez ok. 20 tys. mieszkańców i odbiera ścieki komunalne ze średnim udziałem ścieków przemysłowych.

Do szczegółowej analizy zostały wybrane ładunki BZT_5 w dopływie do oczyszczalni. Charakterystykę zebranych danych przedstawiono w tab. 2. Celem prowadzonych badań było wyznaczenie dla tych zmiennych parametru P_{85} (85. percentyl) z jak najmniejszym błędem.

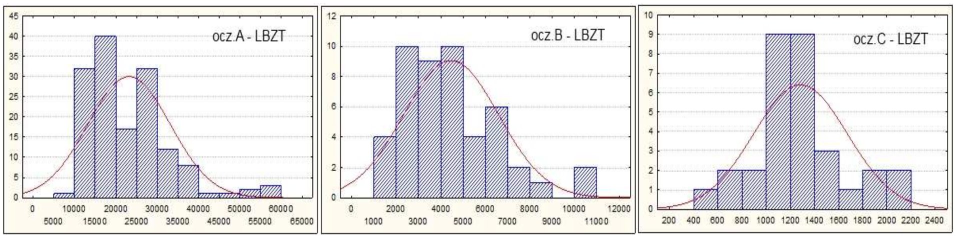
Wstępna ocena statystyczna zebranych danych pokazuje, że dopływy do badanych oczyszczalni nie są zgodne z rozkładem normalnym. Świadczą o tym przede wszystkim:

- rozbieżność pomiędzy wartością średnią a medianą,
- dodatnia skośność świadcząca o asymetrii rozkładu empirycznego (dla rozkładu normalnego skośność wynosi 0) – w badanych przypadkach histogramy rozkładów empirycznych są lewostronnie asymetryczne,
- dodatnia kurtoza świadcząca o bardziej stromym przebiegu histogramu empirycznego w stosunku do rozkładu normalnego (dla którego kurtoza wynosi 0).

Tabela 2. Statystyki opisowe dla ładunku BZT_5 w dopływie do badanych oczyszczalni ścieków

Oczyszczalnia / parametr	***A***	***B***	***C***
Liczba obsługiwanych mieszkańców	300,000	35,000	16,000
Obciążenie oczyszczalni (*RLM*)	385,000	74,000	21,300
Liczebność próby LBZTdop (*n*)	149	48	31
Średnia z próby (m_x) [kg/d]	23063	4437	1279
Mediana	20229	4104	1257
Odchylenie standardowe (s_x)	9906	2115	386
Współczynnik zmienności (v_x)	0,43	0,48	0,30
Skośność	1,300	1,075	0,619
Kurtoza	2,158	1,124	0,661
Percentyl 85 (P_{85})	31941	6448	1641
Błąd oszacowania parametrów rozkładu	3246	1221	277

Podobne wnioski można wyciągnąć, analizując kształt histogramów empirycznych, na które naniesiono linie hipotetycznego rozkładu normalnego (rys. 3.). Kształt tych linii znacznie odbiega od przebiegu histogramów, co potwierdza brak normalności analizowanych zmiennych losowych. Histogramy pokazane na rys. 3. wskazują raczej na logarytmiczno-normalny przebieg zmienności parametru LBZTdop, co jest zresztą zgodne z wynikami przedstawionymi w innych publikacjach [1, 6].



Rys. 3. Histogramy zmiennych empirycznych LBZT dla oczyszczalni *A*, *B*, *C*

Dodatkowo, aby zweryfikować dokładność oszacowania parametru P_{85} za pomocą zaproponowanych metod (empiryczna, aproksymacja rozkładem normalnym lub log-normalnym), przeprowadzono symulację zmian ładunków BZT w dopływie do badanych oczyszczalni z wykorzystaniem metody Monte-Carlo. Polega ona na wielokrotnym przeliczaniu deterministycznego modelu z wykorzystaniem „niepewnych” danych wejściowych, uzyskiwanych z generatora liczb pseudolosowych za pomocą jednego ze znanych teoretycznych rozkładów statystycznych, dopasowanego do rzeczywistych wartości tych danych. W tym przypadku wykorzystano arkusz kalkulacyjny, w którym zastosowano formuły generujące liczby losowe z zadanych przedziałów, zgodnie z prawdopodobieństwem ich występowania, określonym w histogramie empirycznym danej zmiennej (dla zapewnienia zgodności symulowanych wartości z rzeczywistym rozkładem prawdopodobieństwa). W ten sposób generowano zestawy 365 wartości każdej zmiennej, wyznaczając dla każdego zestawu parametr P_{85}. Po przeprowadzeniu 100 kolejnych symulacji wyznaczono wartość średnią 85. percentyla dla każdego badanego obiektu.

Ostatnim etapem badań było zestawienie i porównanie uzyskanych wyników (tab. 3.). Na podstawie uzyskanych rezultatów można zauważyć, że wszystkie analizowane metody wyznaczania parametrów charakterystycznych rozkładu zmiennej losowej (w szczególności percentyla P_{85}) określają wartości estymowanych parametrów z wystarczającą dla praktyki inżynierskiej dokładnością (błąd względny nie przekracza 10% dla dowolnej analizowanej metody). Należy podkreślić, że dobrą zgodność parametrów estymowanych otrzymano za pomocą rozkładu empirycznego, ze wzorcowym rozkładem uzyskanym z symulacji danych metodą Monte-Carlo. Wynika stąd, że przy liczbie pomiarów nie mniejszej niż 30 rozkład empiryczny może być najlepszym i najprostszym sposobem wyznaczania parametrów, jakie należy przyjąć do obliczeń technologicznych oczyszczalni ścieków.

Tabela 3. Porównanie estymatorów percentyla P_{85} dla ładunku BZT_5 w dopływie do badanych oczyszczalni ścieków

Oczyszczalnia / parametr		***A***	***B***	***C***
Symulacja Monte-Carlo	wartość P_{85} [kg/d]	31165	6636	1673
	błąd względny [%]*	-	-	-
Rozkład empiryczny	wartość P_{85} [kg/d]	31941	6448	1641
	błąd względny [%]*	2,5	–2,8	–1,9
Rozkład normalny (rys. 1.)	wartość P_{85} [kg/d]	33425	6622	1682
	błąd względny [%]*	7,2	–0,2	0,5
Rozkład log-normalny (rys. 2.)	wartość P_{85} [kg/d]	32483	6430	1661
	błąd względny [%]*	4,2	–3,1	–0,7

* błąd względny wyznaczono w stosunku do wartości P_{85} z symulacji

Z kolei metody wykorzystujące rozkład teoretyczny jako podstawę oszacowania parametrów rozkładu zmiennej znajdą zastosowanie przy małej liczbie dostępnych danych. Co prawda, zgodnie z centralnym twierdzeniem granicznym, wzrost liczebności zbioru danych powinien dawać coraz lepszą zgodność z rozkładem normalnym, jednak nie znalazło ono potwierdzenia w przeprowadzonych analizach (błąd względny oszacowania parametru P_{85} jest największy dla zbioru danych o największej liczebności). Może to być jednak spowodowane błędem oszacowania wartości średniej i odchylenia standardowego pochodzących z rozkładu empirycznego, które decydują o wartościach odczytanych z nomogramów, co wymaga dalszych badań i analiz.

4. Podsumowanie

Przedstawione w artykule metody statystyczne badania danych wyjściowych, jakie są podstawą przyjęcia określonych wartości parametrów projektowych, powinny wejść na stałe do procedur stosowanych w obliczeniach inżynierskich. Należy przy tym dążyć do ograniczenia do minimum stosowania wskaźników jednostkowych, które mogą opisywać jedynie średnie warunki pracy oczyszczalni. Natomiast nie powinny być one stosowane do wymiarowania obiektów technologicznych. Uzasadnione jest też prowadzenie dalszych badań nad opracowaniem optymalnej strategii oszacowania parametrów projektowych, w zależności od liczebności oraz właściwości statystycznych wyjściowego zbioru danych. Szczególnie przydatne w tych badaniach mogą być modele symulacyjne, takie jak np. wykorzystana w niniejszej pracy metoda Monte-Carlo.

Literatura

1. ATV-Regelverk-Abbwasser-Arbeitsblatt A 131: Bemessung von einstufigen Belebungsanlagen ab 5000 Einwohnerwerten, 1991.
2. Bever J., Stein A., Teichman A.: Zaawansowane metody oczyszczania ścieków, Wydaw. Projprzem-EKO, Bydgoszcz 1997, s. 163-173.
3. Devore J.L.: Probability and statistics for engineering and sciences, Brooks/Cole Publ. Co., Pacific Grave, California 1991, s. 241.

APPLICATION OF STATISTICAL METHODS IN DESIGN OF WASTEWATER TREATMENT PLANTS

A b s t r a c t

In the process of designing wastewater treatment plant engineer very often has to deal with decision-making problems, starting from selection of reliable computational parameters for technological calculations and ending with accepance of optimum designing variant. In the paper,

various statistical methods that can be used in the analysis and evaluation of preliminary data describing inflow to the wastewater treatment plant are presented. Aim of this analysis is to estimate the most accurate design parameters. For the verification of research results Monte-Carlo simulation was used. Presented work is part of research grant S/WBiIŚ/22/08 from Bialystok University of Technology.

Złożono w Oficynie Wydawniczej w lipcu 2011 r.