

Uczenie maszynowe w prognozowaniu ilości ścieków dla oczyszczalni ścieków miasta Rzeszowa na podstawie warunków pogodowych

Bartosz Kowal, Patryk Organiściak, Paweł Kuraś

Politechnika Rzeszowska im. Ignacego Łukasiewicza, Wydział Elektrotechniki i Informatyki, ul. Marii Skłodowskiej-Curie 8/2, 35-037 Rzeszów

Adam Masłoń

Politechnika Rzeszowska im. Ignacego Łukasiewicza, Wydział Budownictwa, Inżynierii Środowiska i Architektury, ul. Powstańców Warszawy 6, 35-959 Rzeszów

Bartosz Wadiak

Miejskie Przedsiębiorstwo Wodociągów i Kanalizacji Sp. z o.o. w Rzeszowie, ul. Naruszewicza 18, 35-055 Rzeszów

Streszczenie: Parametry oczyszczalni ścieków w Polsce muszą spełniać określone normy, które regulują jakość odpływu po procesie oczyszczania. Ilość oraz jakość dopływających ścieków zależy od wielu czynników, między innymi od warunków pogodowych. Prognozowanie tych parametrów pozwala zapewnić optymalną pracę oczyszczalni, co przyczynia się do redukcji kosztów ich pracy. W tym celu, korzystając z danych pogodowych, przeprowadzono próbę oszacowania ilości ścieków dopływających do oczyszczalni w Rzeszowie. Wykorzystano ponad 1000 modeli uczenia maszynowego (ML), w tym także modele statystyczne, takie jak ARIMA i SARIMAX, oraz algorytmy ML, takie jak KNN i sieci neuronowe, w różnych konfiguracjach i przedziałach czasowych. Uzyskano najmniejszy średni błąd bezwzględny (MAE) na poziomie 3598 m³ oraz błąd średniokwadratowy (RMSE) równe 4808 m³. Badanie pokazało, jak wybór parametrów oraz różnych typów modeli predykcyjnych (statycznych, dynamicznych, uczenia maszynowego) wpływa na dokładność prognoz, co bazując wyłącznie na podstawowych danych czasowych, okazuje się być wymagającym procesem.

Słowa kluczowe: oczyszczalnie ścieków, sieci neuronowe, systemy sterowania, uczenie maszynowe

1. Wprowadzenie

Woda jest jednym z najważniejszych zasobów naturalnych, a jej jakość i dostępność są kluczowe dla zachowania zdrowia ludzi i równowagi ekosystemów [1, 2]. Ochrona i utrzymanie jakości wód, w tym proces oczyszczania ścieków, stają się coraz większym wyzwaniem, zwłaszcza w kontekście zmieniających się warunków klimatycznych i rosnącego nacisku na zrównoważony rozwój. Odnawialna słodka woda, choć stanowi niewielką część światowego zaopatrzenia w wodę, jest niezbędna dla ekosystemów, wspierając działalność człowieka (m.in. woda pitna, nawadnianie i zastosowania przemysłowe). Zmiany klimatu i rosnąca nierównowaga między zużyciem,

a podażą wody, będą miały znaczący wpływ na jej obieg. Skutkiem tego może okazać się spadek dostępności słodkiej wody na mieszkańca i zagrożenie ekosystemów wodnych [3]. Sprawna gospodarka wodno-ściekowa może mieć duże znaczenie dla środowiska, a ponadto sam proces oczyszczania ścieków stanowi kluczowy krok w kierunku poprawy jakości środowiska wodnego [4]. Skład i ilość ścieków ma również istotny wpływ na inne czynniki środowiskowe, m.in. na produkcję biogazu w celu samowystarczalności energetycznej oczyszczalni [5].

Prędkość i kierunek przemieszczania się deszczu wpływają na ilość powstających ścieków w sieci kanalizacyjnej [6], co ma istotne znaczenie dla formowania się ilości i jakości ścieków dopływających do oczyszczalni. Opady deszczu, temperatura oraz zmiany sezonowe oddziałują na rodzaj i ilość substancji chemicznych i zanieczyszczeń biologicznych, jak również na fizyczne właściwości ścieków. Zrozumienie tych zależności jest możliwe [7] i ma duże znaczenie dla zapewnienia efektywnej pracy oczyszczalni ścieków, a zatem również redukcji kosztów związanych z ich funkcjonowaniem.

Celem tej pracy jest zrozumienie zależności między warunkami pogodowymi a parametrami ścieków, potwierdzenie korelacji między tymi danymi, oraz opracowanie modeli pro-

Autor korespondujący:

Paweł Organiściak, org@prz.edu.pl

Artykuł recenzowany

nadesłany 18.02.2024 r., przyjęty do druku 10.01.2025 r.



Zezwala się na korzystanie z artykułu na warunkach licencji Creative Commons Uznanie autorstwa 4.0 Int.

gnozujących na ich podstawie. Efekty badania mogą pomóc w zbilansowanym, nowoczesnym i efektywnym zarządzaniu oczyszczalniami ścieków, jak i również wskazać kierunek dla dalszych badań.

1.1. Kontekst i znaczenie badań

Rozwinięcie znaczenia ochrony środowiska jest istotnym tematem, który znajduje odzwierciedlenie w licznych źródłach. Takie zagadnienia, jak koncepcja zerowej ilości odpadów (ang. *The Concept of Zero Waste* [8]) czy też ograniczanie produkcji plastiku ciągle zyskują na popularności. Skuteczne gospodarowanie odpadami i niezawodna praca oczyszczalni ścieków mają znaczny wpływ na przyrodę biorąc pod uwagę, że oczyszczalnie ścieków zostały zidentyfikowane jako ważne źródła uwalniania tworzyw sztucznych do środowiska wodnego i lądowego, co w konsekwencji może prowadzić do dalszego zanieczyszczenia [9].

Nierównomierność dopływających ścieków, jak również towarzyszące im często zmiany jakościowe stanowią zasadniczy czynnik powodujący zakłócenia w prawidłowym przebiegu procesów oczyszczania ścieków. Bilans ilościowo-jakościowy ścieków dopływających do oczyszczalni jest źródłem cennych informacji zarówno dla projektanta, jak i dla eksploatatora obiektu – rzetelna wiedza na temat dynamiki dopływającego ładunku zanieczyszczeń do oczyszczalni wpływa na poprawną eksploatację zastosowanej technologii oczyszczania ścieków [10].

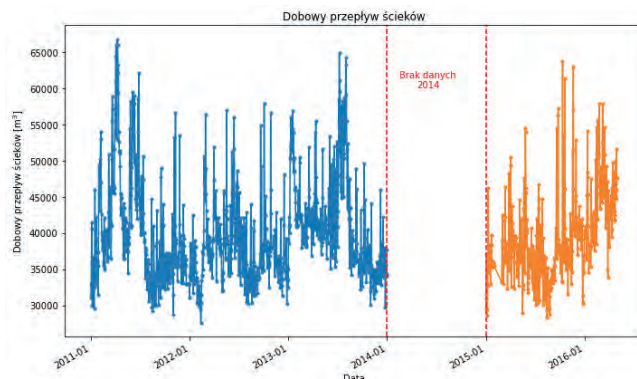
Niniejsze badanie może przyczynić się zarówno do poprawy efektywności procesów oczyszczania ścieków oraz zmniejszenia negatywnego wpływu na środowisko naturalne. Innym istotnym aspektem badań są wymagania prawne stawiane oczyszczalniom i idące za tym ryzyko kar. Państwo przewiduje kary pieniężne za przekraczanie norm [32]. Aby temu przeciwdziałać opracowano metody do wspierania podejmowanych decyzji [11] oraz modelowania ryzyka w eksploatacji oczyszczalni ścieków [12, 13].

2. Oczyszczalnia ścieków w Rzeszowie

Cykliczne modernizacje oraz rozbudowa oczyszczalni ścieków w Rzeszowie, zrealizowane w różnych okresach (1992–1998, 2001, 2003–2004, 2013–2014), przyczyniły się do ewolucji jej charakterystyki technologicznej. Zmiany wprowadzone podczas tych procesów mogą wymagać regularnych aktualizacji badań, aby monitorować aktualny stan oczyszczalni. Oczyszczalnia jest zlokalizowana przy ul. Ciepłowniczej 2, w północnej części miasta, na prawym brzegu rzeki Wisłok [14].

2.1. Parametry jakości wody w oczyszczalniach ścieków

Dane dotyczące jakości i ilości ścieków, który wykorzystano w analizie, pochodzą z oczyszczalni ścieków zlokalizowanej w mieście Rzeszów, zaś dane meteorologiczne zostały pozyskane z publicznie dostępnych danych ze stacji pogodowej działającej w obszarze Rzeszowa i Jasionki [33]. Zbiór danych w charakterze dobowym jest dostępny od 1 stycznia 2011 r. do 30 kwietnia 2016 r. z wyłączeniem okresu, w którym oczyszczalnia był modernizowana (rys. 1). Dane, które zawierają ładunek ścieków można opisać za pomocą kryteriów ilościowych, jak i jakościowych (tab. 1), a ich charakterystykę opisano w tab. 2.



Rys. 1. Dobowy przepływ danych w oczyszczalni ścieków Rzeszów
Fig. 1. Daily data flow at Rzeszow wastewater treatment plant

Tab. 1. Badane parametry ścieków surowych
Tab. 1. Raw wastewater parameters analyzed

Parametr	Opis
Z _{og}	Zawiesina ogólna (g/m ³)
BZT ₅	Biochemiczne zapotrzebowanie na tlen (gO ₂ /m ³)
ChZT	Chemiczne zapotrzebowanie na tlen (g/m ³) [15]
N _{og}	Azot ogólny (g/m ³)
NH ₄ -N	Azot amonowy (g/m ³)
Q _{dśr}	Przepływ średniodobowy (m ³ /d) [16]

Tab. 2. Charakterystyka danych dotyczących przepływu ścieków surowych w Rzeszowie od 1 stycznia 2011 r. do 30 kwietnia 2016 r. z wyłączeniem okresu modernizacji (rok 2014)

Tab. 2. Characteristics of raw wastewater data in Rzeszow from 1st January 2011 to 30 April 2016 excluding the modernization period (2014 year)

Parametr	Dni	Średnia	Std	Min	25,00 %	50,00 %	75,00 %	Max
Z _{og}	1554	416,63	102,81	80,00	362,00	418,00	470,00	1200,00
BZT ₅	1554	362,91	91,81	38,00	310,00	368,00	417,00	1290,00
ChZT	1554	902,70	216,40	159,00	790,00	900,00	1000,00	2510,00
N _{og}	1554	68,29	12,59	10,00	62,00	70,00	76,00	146,00
NH ₄ -N	1554	52,61	10,85	1,00	47,00	54,00	60,00	97,00
Q _{dśr}	1554	39 820,99	6524,76	27 610,00	35 428,00	38 276,50	42 451,25	66 773,00

2.2. Korelacja między opadami a ilością ścieków

Badania dotyczące ścieków jednoznacznie sugerują, że zachodzi korelacja między warunkami pogodowymi a jakością i ilością dopływających do oczyszczalni ścieków oraz, że jest ona zależna od przesunięcia sygnałów w czasie [17, 18]. Ilość ścieków dopływających do oczyszczalni w Rzeszowie w ciągu jednej doby może być przewidywana za pomocą sztucznych sieci neuronowych. W procesie prognozowania ilości ścieków, najważniejszymi czynnikami okazują się: wysokość opadów, ilość ścieków zmierzona z opóźnieniem jednego i dwóch dni oraz stan wody w odbiorniku, dostarczony z opóźnieniem wynoszącym jeden dzień [19]. Bartkiewicz i autorzy ustalili, że możliwe jest zbliżone prognozowanie wartości parametrów ilościowych ścieków z użyciem sieci neuronowych, bazując na parametrach, takich jak dopływ ścieków zmierzony z dwóch poprzednich dób oraz dopływ ścieków z trzech poprzednich dób. Jest to jedno z podejść do predykcji ilościowej, natomiast w tej pracy skupiono się na predykcji za pomocą wyłącznie danych czasowych, aby ocenić ich skuteczność w predykcji dobowej ilości ścieków [19].

3. Metodyka badań

Jednym z wyzwań w prognozowaniu danych metodami uczenia maszynowego ML (ang. *Machine learning*) [20] jest wybór odpowiednich parametrów konfiguracyjnych. W niniejszym rozdziale analizowano możliwości predykcji dobowego przepływu ścieków do oczyszczalni w Rzeszowie, wykorzystując metody ML w różnych konfiguracjach. Analizowane w badaniu dane składają się z informacji pogodowych, takich jak opady i temperatury oraz dobowy dopływ ścieków. Zbiór danych uzupełniono o dni tygodnia oraz miesiące, aby uwidatnić zależności między poszczególnymi dniami. Zbiór badano w konfiguracji danych stosując dni tygodnia z miesiącami. Dodatkowo badania te uwzględniają trzy różne metody opadów w ujęciu czasowym.

W analizie wykorzystano dwie główne kategorie metod: modele statystyczne (ARIMA, SARIMAX) oraz algorytmy uczenia maszynowego (drzewa decyzyjne, KNN, MLP, las losowy). Warto zaznaczyć, że nie wszystkie z tych metod są algorytmami uczenia maszynowego, jednak wszystkie służą do predykcji danych, których precyzję zbadano w trakcie analizy.

Dane pogodowe zaczerpnięte ze stacji meteorologicznej znajdującej się w Jasionce mają szereg parametrów, takich jak temperatura powietrza, wielkość opadów i ich czas trwania. Biorąc pod uwagę, że wody opadowe mogą dopływać do oczyszczalni z pewnym opóźnieniem czasowym o charakterze falowym [21, 22], rozważono je w trzech wariantach:

1. t_1 – dobowy sumaryczny opad deszczu zarejestrowany w dniu pomiaru przepływu w oczyszczalni,
2. t_2 – dobowy sumaryczny opad rejestrowany od 12:00 dnia poprzedniego do 12:00 dnia pomiarowego w oczyszczalni,
3. t_3 – dobowy sumaryczny opad rejestrowany dnia poprzedzającego dzień pomiaru w oczyszczalni.

Proponowany podział umożliwił sprawdzenie wpływu opadów na precyzję budowanych modeli. W ocenie skuteczności modeli wzięto pod uwagę trzy wskaźniki:

1. $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$ – tzw. średni błąd bezwzględny [23].

Jest to średnia wartość bezwzględnych różnic między prognozowanymi a rzeczywistymi wartościami. MAE dostarcza informacji, jak daleko od rzeczywistych wartości są prognozy.

2. $Max Error = \max(|y_i - \hat{y}_i|)$ – to największa różnica między prognozowaną a rzeczywistą wartością w badanych danych.

Ta statystyka pokazuje, jak duży może być najgorszy przypadek błędu w prognozach.

$Tolerance - Bounded Accuracy =$

$$= \frac{\sum_{i=1}^n 1(|y_i - \hat{y}_i| < Tolerancja \times |y_i|)}{n} \times 100 \%,$$

gdzie: y_i = rzeczywista wartość i -tej obserwacji, \hat{y}_i = przewidywana wartość i -tej obserwacji, Tolerancja = ustalony próg tolerancji błędu, np. 10 %, n = liczba obserwacji.

Ostatni z modeli poprawności prognoz został zdefiniowany w kontekście niniejszych badań. Dokładność obliczana jest jako procent prognoz, które są wystarczająco blisko rzeczywistych wartości. Błędy rzędu 10 % do danych rzeczywistych były traktowane jako poprawne, pozostałe jako dane błędne.

4. Prognozowanie parametrów oczyszczalni na podstawie warunków pogodowych

Pierwszą ze stosowanych w badaniu metod do analizy szeregu czasowego jest analiza wykładnika Hursta [24] do analizy zależności długoterminowej. Hurst w swojej analizie podczas projektu budowy tamy na rzece Nil odkrył, że poziom rzeki Nil nie jest procesem losowym. Wykładnik ten jest narzędziem statystycznym służącym do klasyfikacji szeregów czasowych na trzy grupy: losowe, gdy $H = 0,5$; antypersystentne, gdy $0 < H < 0,5$ oraz persystentne, gdy $0,5 < H < 1$. Szacowanie wykładnika Hursta w swojej podstawowej wersji związane jest z analizą $R \setminus S$ czyli tzw. analizy przeskalowanego zakresu. Obliczony wykładnik Hursta dla danych przepływu ścieku wynosi 0,736, co wskazuje na zależności długoterminowe, czyli wzorce przepływu ścieków utrzymują się przez dłuższy czas. Gdyby zależności były krótkoterminowe lub losowe, oznaczałoby to, że przepływy szybko się zmieniają i byłyby trudniejsze do późniejszego przewidywania i analiz statystycznych.

4.1. Analiza stacjonarności

Stacjonarność jest właściwością szeregów czasowych, w której ich statystyczne charakterystyki, takie jak średnia, wariancja i autokorelacja, nie zmieniają się w czasie. Metoda badania szeregów czasowych pod względem stacjonarności jest jednym z podstawowych metod do analizy danych ekonomicznych [25]. Badanie stacjonarności w przypadku dobowego przepływu ścieków jest istotną kwestią, ponieważ pozwala z matematycznego punktu widzenia ocenić, czy te właściwości statystyczne szeregu czasowego pozostają stałe, co może rzutować na wiarygodności modeli ponieważ wiele metod statystycznych i ekonometrycznych zakłada stacjonarność danych. Zrozumienie, czy dane przepływu są stacjonarne, jest również ważne do identyfikacji trendów i sezonowości, co ma istotne implikacje dla planowania, projektowania i zarządzania infrastrukturą kanalizacyjną.

Na podstawie podanych wyników przedstawionych w tab. 3 można zauważyć, że testy ADF (ang. *Augmented Dickey-Fuller*) i PP (Phillips-Perron) sugerują stacjonarność badanego szeregu czasowego, ponieważ ich wartości p (P-value) są bardzo niskie – 0,01, co wskazuje na odrzucenie hipotezy o niestacjonarności. Test KPSS (Kwiatkowski-Phillips-Schmidt-Shin) daje wyższą wartość $p = 0,079$, co może sugerować stacjonarność, ale z mniejszą pewnością. Wartości ERS (Elliott-Rothenberg-Stock) wydają się mieć mieszane wyniki z jedną wysoką wartością P-test na poziomie 0,97 i innymi wartościami krytycznymi, co może wska-

Tab. 3. Wyniki stacjonarności dla dobowego przepływu ścieków

Tab. 3. Stationarity results for daily wastewater flow

	ADF	KPSS	PP	ERS
P-wartość/P-test	0,01	0,079	0,01	0,97 3,26 4,48
Stacjonarność	Stacjonarny	Stacjonarny	Stacjonarny	Stacjonarny

zywać niepewności w stacjonarności szeregu. Użycie tych metod może pozwolić na określenie założenia, że wartości dobowego przepływu ścieków wskazują na potencjalną stacjonarność szeregu czasowego. Zapewnia to, że tworzone modele będą pracować na danych o stabilnych właściwościach statystycznych, co potencjalnie zwiększa ich zdolność do generowania dokładniejszych, a także bardziej wiarygodnych prognoz dobowego przepływu ścieków za pomocą modeli uczenia maszynowego oraz modeli statystycznych.

4.2. Analiza i porównanie metod predykcji

Pierwszą użytą metodą do predykcji była ARIMA [26], która opiera się na statystycznej koncepcji korelacji szeregowej, w której przeszłe punkty danych wpływają na przyszłe punkty danych. Jest to model, który przybiera tylko jedno wejście, tj. szereg czasowy i jedno wyjście, którym jest dobowy przepływ ścieków. Podczas badań stworzono 75 różnych konfiguracji modelu ARIMA, gdzie zmieniane były trzy parametry:

autoregresji, różnicowania i średniej ruchomej, które mogą odgrywać kluczową rolę w prognozowaniu szeregów czasowych.

Analizując przedstawione w tab. 4 wyniki modelu ARIMA, zaobserwowano, że wartości MAE wahały się od 3800,88 do 162 433,32, RSME od 5306,39 do 187 392,46, zaś stopień poprawności modelu nie przekroczył 64 %, co wskazuje, że aż 36 % prognoz było niepoprawnych. Warto też zwrócić uwagę na znaczącą różnicę w maksymalnych wartościach między danymi rzeczywistymi a realnymi.

4.3. Drzewo decyzyjne

W tworzeniu i ocenie poprawności predykcji w tych testowano różne konfiguracje drzew decyzyjnych [27], zmieniając parametry, takie jak głębokość drzewa (od 10 do 100), liczbę próbek branych pod uwagę przy podziale w węzle (od 2 do 30) oraz kryteria podziału uwzględniając trzy różne czasy opadów. Łącznie dla każdego z czasu opadów t przebadano po 378 modeli. W tab. 5 przedstawiono wyniki z ponad 1000 modeli

Tab. 4. Wyniki dla modeli ARIMA

Tab. 4. Results for ARIMA models

	Min	Max	Średnia	Mediana
MAE	3800,88	162 433,32	9764,63	4619,38
RMSE	5306,39	187 392,46	28 799,66	5968,63
Max Error	20 298,25	329 831,78	34 662,86	23 797,53
Accuracy	0,64 %	63,38 %	50,68 %	53,32 %

Tab. 5. Wyniki dla modeli drzewa decyzyjnego z różnymi czasami opadów

Tab. 5. Results for Decision Tree models with different rainfall times

		Min	Max	Średnia	Mediana
t_1	MAE	4399,46	5125,64	4588,75	4523,10
	RMSE	5947,42	7119,09	6202,14	6106,81
	Max Error	28 167,60	29 011	28 765,55	28 857,5
	Accuracy	51,17 %	55,46 %	53,67 %	53,53 %
t_2	MAE	4389,76	5327,36	4569,26	4477,13
	RMSE	6069,4	7411,96	6310,61	6201
	Max Error	27 967,28	33 413,00	28 903,29	28 906,18
	Accuracy	51,17 %	56,95 %	55,28 %	55,67 %
t_3	MAE	4472,14	5662,96	4751,08	4632,12
	RMSE	6111,1	7779,01	6423,28	6301,46
	Max Error	28 750,00	29 336,84	29 062,98	29 336,84
	Accuracy	46,25 %	52,67 %	50,81 %	51,17 %

drzewa decyzyjnego. W tej metodzie uzyskano najniższy błąd MAE rzędu 4400 m³, RMSE 6100 m³, przy współczynniku poprawności na poziomie 56,95 % dla t_2 .

4.4. KNN

W badaniach rozważono również zastosowanie modelu K najbliższych sąsiadów [28]. W testach przebadano różne konfiguracje modelu, zmieniając parametry, takie jak liczba sąsiadów (K : 16, 32, 64, 128, 256, 512, 1024), metrykę odległości (euklidesowa, manhattan) oraz algorytmy wyboru sąsiadów. Łącznie dla każdego z czasu opadów t przebadano po 14 modeli KNN. W ponad 1000 stworzonych modeli KNN można zaobserwować poziom wyników zbliżony do wyników drzewa decyzyjnego. Wyniki zaprezentowano w tab. 6. W tej metodzie uzyskano

najmniejsze błędy MAE rzędu 4200 m³ oraz RMSE 5800 m³, przy współczynniku poprawności na poziomie 56,31 % dla t_1 .

4.5 MLP

W analizie predykcji zastosowano również model perceptronu wielowarstwowego (MLP) [29]. Test obejmował różne konfiguracje modelu, takie jak liczba neuronów w warstwach ukrytych (16, 32, 64, 128, 256, 512, 1024, 2048), liczba epok treningowych (100, 200), a także dodawanie ukrytej warstwy drugiej (64, 128, 256, 512, 1024) i ukrytej warstwy trzeciej, której liczba neuronów zależała od warstwy drugiej. Łącznie dla każdego z czasu opadów przebadano po 100 modeli MLP. Wyniki tych modeli przedstawiono w tab. 7. Ta metoda zwróciła najwyższy, jak dotąd wskaźnik poprawności wynoszący

Tab. 6. Wyniki dla modeli KNN z różnymi czasami opadów

Tab. 6. Results for KNN models with different rainfall times

		Min	Max	Średnia	Mediana
t_1	MAE	4212,59	4819,78	4388,69	4322,88
	RMSE	5769,45	6293,71	5930,28	5855,44
	Max Error	25 747,84	27 553,21	26 483,74	26 307,56
	Accuracy	47,53 %	56,31 %	53,48 %	54,60 %
t_2	MAE	4280,39	4821,04	4438,07	4364,09
	RMSE	5872,27	6296,31	6007,99	5952,57
	Max Error	25 697,23	27 514,51	26 406,29	26 097,86
	Accuracy	47,96 %	55,46 %	52,90 %	53,96 %
t_3	MAE	4363,75	4866,41	4494,09	4405,66
	RMSE	5971,32	6320,69	6064,36	6026,84
	Max Error	25 075,32	27 222,48	26 222,06	25 925,29
	Accuracy	47,10 %	56,31 %	52,53 %	53,10 %

Tab. 7. Wyniki dla modeli MLP z różnymi czasami opadów

Tab. 7. Results for MLP models with different rainfall times

		Min	Max	Średnia	Mediana
t_1	MAE	4055,96	5402,65	4392,93	4349,71
	RMSE	5736,09	6507,79	5939,9	5920,22
	Max Error	23 638,73	29 147,45	26 193,86	26 152,49
	Accuracy	37,9 %	62,95 %	54,8 %	55,24 %
t_2	MAE	4106,88	4983,03	4427,06	4377,84
	RMSE	5847,58	6380,89	6028,12	6002,11
	Max Error	24 298,83	28 601,79	26 022,14	25 935,07
	Accuracy	43,68 %	64,66 %	54,37 %	54,60 %
t_3	MAE	4096,14	5709,14	4514,20	4467,99
	RMSE	5902,29	7043,86	6121,34	6097,94
	Max Error	23 300,21	29 177,9	25 688,99	25 607,96
	Accuracy	39,4 %	64,66 %	53,44 %	53,42 %

Tab. 8. Wyniki dla modeli lasu losowego z różnymi czasami opadów

Tab. 8. Results for random forest models with different rainfall times

		Min	Max	Średnia	Mediana
t_1	MAE	4569,32	4732,98	4620,28	4589,40
	RMSE	6231,96	6433,56	6288,73	6243,57
	Max Error	29 068,40	30 227,68	29 529,07	29 410,11
	Accuracy	53,31 %	55,24 %	54,33 %	54,38 %
t_2	MAE	4603,12	4665,70	4622,97	4611,53
	RMSE	6439,29	6458	6449,28	6449,91
	Max Error	29 563,20	30 440,02	30 023,92	30 046,22
	Accuracy	52,03 %	55,03 %	54,06 %	54,60 %
t_3	MAE	4927,91	5042,70	4967,38	4949,45
	RMSE	6588,11	6811,1	6665,16	6629,6
	Max Error	29 014,55	30 123,76	29 437,04	29 304,92
	Accuracy	48,39 %	49,89 %	49,19 %	49,25 %

64,66 % poprawności z błędem MAE rzędu 4100 m³ oraz RMSE 5800 m³ dla t_2 oraz t_3 . Niestety model też zwraca maksymalną różnicę błędu na poziomie 23 000 m³, co stanowi aż około 58 % realnego przepływu ścieków.

4.6. Random Forest (las losowy)

Las losowy jest jedną z bardziej popularnych metod ML, szeroko stosowaną w analizie danych [30]. W przeprowadzonym teście istotnym parametrem była liczba drzew decyzyjnych (estymatorów) w lesie. Analiza polegała na zmianie liczby drzew (10, 50, 100, 200) dla każdego z modeli. Łącznie, dla każdego okresu opadów, przebadano po cztery modele Random Forest. Algorytm Random Forest charakteryzował się niewielkimi zmianami w samej strukturze modelu. Jak przedstawiono w tab. 8, wyniki dla MAE, podobnie jak w przy-

padku wcześniejszych modeli ML, wynosiły około 4600 m³, a RMSE około 6200 m³. Najskuteczniejsze modele Random Forest charakteryzowały się dokładnością na poziomie 55,24 % dla t_1 .

4.7. SARIMAX

W rozdziale 4.2 stosowano modele ARIMA, które nie pozwalały na uwzględnienie zmiennych zewnętrznych, m.in. opadów, dlatego w kolejnych seriach obliczeń użyto modelu SARIMAX. SARIMAX to zaawansowany model statystyczny służący do analizy i prognozowania szeregów czasowych, który uwzględnia zarówno sezonowość, jak i wpływ zewnętrznych zmiennych regresorów [31]. Eksperymenty te obejmowały testowanie różnych konfiguracji modelu SARIMAX, zmieniając parametry autoregresji (od 0 do 4), stopień różnicowania (od 0 do 2)

Tab. 9. Wyniki dla modeli SARIMAX z różnymi czasami opadów

Tab. 9. Results for SARIMAX models with different rainfall times

		Min	Max	Średnia	Mediana
t_1	MAE	3598,02	157492,53	16069,43	4805,26
	RMSE	4908,4	181754,75	33045,22	6549,72
	Max Error	17 337,93	319 820,90	41 363,50	25 612,39
	Accuracy	0 %	65,31 %	39,76 %	54,17 %
t_2	MAE	3757,94	20 629 330 000	275 072 800	4907,21
	RMSE	5124,77	21 047 340 397	2430 337 529	6779,28
	Max Error	19841,84	28 258 790 000	376 830 400	27,546
	Accuracy	0 %	63,38 %	40,39 %	52,46 %
t_3	MAE	3857,05	200 200,62	17 651,74	5085,47
	RMSE	5283,33	200 408,08	40 448,4	7036,39
	Max Error	20 296,57	331 182,51	50 567,76	27 974,81
	Accuracy	0 %	62,95 %	40,10 %	51,82 %

Tab. 10. Podsumowanie dla skuteczności modeli ML

Tab. 10. Summary for effectiveness of ML models

Model	MAE	RMSE	Max Error	Precyzja	Czas opadów
ARIMA	3800,88	5306,39	20 298,25	63,38 %	brak
Decision tree	4389,76	6069,4	27 967,28	56,95 %	t_2
KNN	4212,59	5769,45	25 075,32	56,31 %	t_1, t_3
MLP	4055,96	5736,09	23 300,21	64,66 %	t_2, t_3
Random forest	4569,32	6231,96	29 014,55	55,24 %	t_1
SARIMAX	3598,02	4908,4	27 974,55	65,31 %	t_1

oraz stopień średniej ruchomej (od 0 do 4). Jako zmienną zewnętrzną ustawiono opady t . Łącznie dla każdego z czasu opadów przebadano po 72 modeli SARIMAX. W tab. 9 zaprezentowano wyniki ze wszystkich stworzonych modeli SARIMAX. Można zaobserwować, że metoda osiągnęła najwyższe wyniki równe 65,31 % z błędem MAE rzędu 3600 m³, RMSE 4900 m³ dla t_1 .

4.8. Analiza wyników i osiągnięty błąd prognoz

W ramach badań przeanalizowano skuteczność sześciu różnych modeli uczenia maszynowego (ARIMA, drzewa decyzyjne, KNN, MLP, las losowy, SARIMAX). Zbiorcze zestawienie wyników uzyskanych z tych metod przedstawiono w tab. 10, która zawiera najbardziej optymalne wyniki dla każdego modelu, w tym wartości MAE, RMSE, maksymalnego błędu oraz dokładność dopasowania danych predykcyjnych do rzeczywistych.

Model SARIMAX osiągnął najniższą wartość błędów MAE wynoszącą 3598,02 m³ oraz RMSE 4908,40 m³ dla wariantu czasu opadów t_1 . Jednocześnie różnica między RMSE a MAE w przypadku tego modelu była najmniejsza spośród wszystkich badanych metod, co potwierdza jego najwyższą jakość dopasowania oraz niewielką wrażliwość na duże odchylenia w prognozach.

Algorytm SARIMAX również uzyskał najwyższy poziom dokładności dopasowania, wynoszący 65,31 % (t_1). Na podstawie uzyskanych wyników można stwierdzić, że SARIMAX jest najbardziej odpowiednim modelem do przyszłych analiz. Jego niski poziom błędu średniego oraz maksymalnego znacząco przewyższa inne modele, które w niektórych przypadkach generowały błędy maksymalne wynoszące nawet 100 % dobowego przepływu ścieków.

Istotnym czynnikiem wpływającym na skuteczność modeli okazał się dobór odpowiedniego zakresu czasowego dla danych opadów, co wskazuje na konieczność uwzględnienia tej zmiennej w przyszłych badaniach.

5. Podsumowanie

Artykuł prezentuje innowacyjną analizę wpływu warunków pogodowych na funkcjonowanie oczyszczalni ścieków, wykorzystując zaawansowane modele predykcyjne oparte na danych meteorologicznych. Ciekawym i innowacyjnym aspektem badań jest uwzględnienie analizy wykładnika Hursta i analizy stacjonarności do analizy zależności długoterminowych oraz wykorzystanie metody Tolerance-Bounded Accuracy w predykcji ścieków, co jest nowatorskim podejściem w porównaniu do tradycyjnej statystyki R². To pozwala na lepsze zrozumienie potencjału modeli w realnych warunkach i ich

praktyczne zastosowanie w prognozowaniu dobowego przepływu ścieków.

Warto podkreślić, że przeprowadzone badania miały bardzo szeroki zakres i obejmowały 12 różnych metod, jednak ze względu na ograniczenia objętości publikacji, przedstawiono tylko najbardziej efektywne z nich. Zastosowane zostały m.in. algorytm najbliższego sąsiada (KNN), ARIMA, drzewa decyzyjne, lasy losowe oraz sieci neuronowe (MLP) dla różnych konfiguracji danych. Mimo że żaden z badanych modeli nie osiągnął optymalnej dokładności, nawet przy założeniu 10 % błędu pomiarowego, ich analiza otwiera drogę do dalszego rozwoju i udoskonalenia metod prognozowania. Badanie wykazało również znaczący wpływ czasu trwania opadów na dokładność modeli, szczególnie w algorytmach Random Forest czy MLP. To odkrycie wskazuje na potrzebę dalszych badań w tym kierunku, z potencjalnym rozszerzeniem bazy danych i dodatkowymi parametrami czasowymi oraz meteorologicznymi, aby jeszcze bardziej poprawić dokładność modeli predykcyjnych.

W długoterminowej perspektywie, uzyskane wyniki stanowią mocne fundamenty do kontynuacji badań nad prognozowaniem dobowego przepływu ścieków. Przyszłe etapy będą włączać nowe dane z kolejnych lat oraz rozważać nie tylko ilościowe, ale i jakościowe aspekty prognozowania przepływu ścieków, co stanowi ogromny potencjał dla rozwoju tej dziedziny.

Bibliografia

- Kiliń Z., *The importance of water and conscious use of water*, "International Journal of Hydrology", Vol. 4, No. 5, 2020, 239–241, DOI: 10.15406/ijh.2020.04.00250.
- Małecki Z.J., Gołębiak P., *Zasoby wodne polski i świata*, „Zeszyty Naukowe. Inżynieria Lądowa i Wodna w Kształtowaniu Środowiska”, Nr 7, 2012, 50–56.
- Jackson R.B., Carpenter S.R., Dahm C.N., et al., *Water in a changing world*, "Ecological applications", Vol. 11, No. 4, 2001, 1027–1045, DOI: 10.1890/1051-0761(2001)011[1027:WIACW]2.0.CO;2.
- Masłoń A., Tomaszek J.A., *Ocena efektywności oczyszczalni ścieków w Lubaczowie*, „Czasopismo Inżynierii Lądowej, Środowiska i Architektury”, Z. 60, Nr 3, 2013, 209–222.
- Organiściak P., Masłoń A.A., Kowal B., et al., *Machine learning-based prediction of biogas production from sludge characteristics in four anaerobic digesters-development of the AD2biogas prediction tool*, "Advances in Science and Technology Research Journal", Vol. 18, No. 8, 2024, 1–15, DOI: 10.12913/22998624/192936.
- Dziopak J., Starzec M., *Wpływ kierunku i prędkości przemieszczania się opadu deszczu na maksymalne szczytowe przepływy ścieków w sieci kanalizacyjnej*, „Czasopismo

- Inżynierii Lądowej, Środowiska i Architektury”, Z. 61, Nr 3, 2014, 63–81.
7. Skoczko I., Ofman P., Szatyłowicz E., *Zastosowanie sztucznych sieci neuronowych do modelowania procesu oczyszczania ścieków w małej oczyszczalni ścieków*, „Rocznik Ochrona Środowiska”, tom 18, cz. 1, 2016, 493–506.
 8. Bogusz M., Matysik-Pejas R., Krasnodębski A., et al., *The concept of zero waste in the context of supporting environmental protection by consumers*, “Energies”, Vol. 14, No. 18, 2021, DOI: 10.3390/en14185964.
 9. Okoffo E.D., O’Brien S., O’Brien J.W., et al., *Wastewater treatment plants as a source of plastics in the environment: a review of occurrence, methods for identification, quantification and fate*, “Environmental Science: Water Research & Technology”, Vol. 5, No. 11, 2019, 1908–1931, DOI: 10.1039/C9EW00428A.
 10. Masłoń A., *Dynamika dopływu ścieków do oczyszczalni w aspekcie funkcjonowania sekwencyjnych reaktorów porcjowych*, „Instal”, Nr 10, 2017, 57–62.
 11. Kuraś P., Strzalka D., Kowal B., et al., *REDUCE – a python module for reducing inconsistency in pairwise comparison matrices*, “Advances in Science and Technology. Research Journal”, Vol. 17, No. 4, 2023, 227–234, DOI: 10.12913/22998624/170187.
 12. Andraka D., Dzieńis L., *Modelowanie ryzyka w eksploatacji oczyszczalni ścieków*, „Rocznik Ochrona Środowiska”, t. 15, cz. 2, 2013, 1111–1125.
 13. Doublein D., Steinhäuser A., *Biogas from waste and renewable resources: an introduction*, John Wiley & Sons, 2011.
 14. Maj K., Masłoń A., *Wieloparametryczna ocena efektywności oczyszczalni ścieków w Rzeszowie przed modernizacją*, „Czasopismo Inżynierii Lądowej, Środowiska i Architektury”, Z. 62, Nr 1, 2015, 299–315, DOI: 10.7862/rb.2015.20.
 15. Sadecka Z., Płuciennik-Koropczuk E., *Frakcje ChZT ścieków w mechaniczno-biologicznej oczyszczalni*, „Rocznik Ochrona Środowiska”, tom 13, 2011, 1157–1172.
 16. Szpindor A., Łomotowski J., *Nowoczesne systemy oczyszczania ścieków*, Wydawnictwo Arkady, Warszawa 2002.
 17. Szeląg B., Studziński J., Chmielowski K., Leśniańska A., Rojek I., *Prognozowanie ilości ścieków dopływających do oczyszczalni za pomocą sztucznych sieci neuronowych z wykorzystaniem liniowej analizy dyskryminacyjnej*, „Ochrona Środowiska”, Vol. 40, No. 4, 2018, 9–14.
 18. Bugajski P., Chmielowski K., Kaczor G., *Wpływ wielkości dopływu wód opadowych na skład ścieków surowych w małym systemie kanalizacyjnym*, „Acta Scientiarum Polonorum. Formatio Circumiectus”, Vol. 15, No. 2, 2016, 3–11, DOI: 10.15576/ASP.FC/2016.15.2.3.
 19. Bartkiewicz L., Szeląg B., Studziński J., *Ocena wpływu zmiennych wejściowych i struktury modelu sztucznej sieci neuronowej na prognozowanie dopływu ścieków komunalnych do oczyszczalni*, „Ochrona Środowiska”, Vol. 38, No. 2, 2016, 29–36.
 20. Zhou Z.-H., *Machine learning*, Springer Nature, 2021.
 21. Bąk Ł., Górski J., Górski K., Szeląg B., *Zawartość zawiesin i metali ciężkich w wybranych falach ścieków deszczowych w zlewni miejskiej*, „Ochrona Środowiska”, Vol. 34, No. 2, 2012, 49–52.
 22. Ociepa E., *Ocena zanieczyszczenia ścieków deszczowych trafiających do systemów kanalizacyjnych*, „Inżynieria i Ochrona Środowiska”, Vol. 14, No. 4, 2011, 357–364.
 23. Chai T., Draxler R.R., *Root mean square error (RMSE) or mean absolute error (MAE)*, “Geoscientific model development discussions”, Vol. 7, 2014, 1525–1534, DOI: 10.5194/gmdd-7-1525-2014.
 24. Hurst H.E., *Long-term storage capacity of reservoirs*, “Transactions of the American Society of Civil Engineers”, Vol. 116, No. 1, 1951, 770–799, DOI: 10.1061/TACEAT.0006518.
 25. Musbah H., Aly H., Little T., *A proposed novel adaptive dc technique for non-stationary data removal*, “Heliyon”, Vol. 9, No. 3, 2023, DOI: 10.1016/j.heliyon.2023.e13903.
 26. Ghysels E., Osborn D.R., Rodrigues P.M., *Chapter 13 Forecasting seasonal time series*, “Handbook of Economic Forecasting”, Vol. 1, 2006, 659–711, DOI: 10.1016/S1574-0706(05)01013-X.
 27. Quinlan J.R., *Induction of decision trees*, “Machine learning”, Vol. 1, 1986, 81–106, DOI: 10.1007/BF00116251.
 28. Fix E., Hodges J.L., *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*, “International Review/Revue Internationale de Statistique”, Vol. 57, No. 3, 1989, 238–247, DOI: 10.2307/1403797.
 29. Taud H., Mas J., *Multilayer perceptron (MLP)*, “Geomatic approaches for modeling land change scenarios”, 2018, 451–455.
 30. Ho T.K., *Random decision forests*, [In:] Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, 278–282, DOI: 10.1109/ICDAR.1995.598994.
 31. Dubey A.K., Kumar A., Garcia-Diaz V., et al., *Study and analysis of SARIMA and LSTM in forecasting time series data*, “Sustainable Energy Technologies and Assessments”, Vol. 47, 2021, DOI: 10.1016/j.seta.2021.101474.

Inne źródła

32. *Rozporządzenie Ministra Gospodarki Morskiej i Żeglugi Śródlądowej z dnia 12 lipca 2019 r. w sprawie substancji szczególnie szkodliwych 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.* (Dz.U. 2019, poz. 1311).
33. Instytut Meteorologii i Gospodarki Wodnej, <https://danepubliczne.imgw.pl>

Machine Learning in Forecasting the Amount of Sewage for the Wastewater Treatment Plant of the City of Rzeszów Based on Weather Conditions

Abstract: The parameters of wastewater treatment plants in Poland must meet certain standards that regulate the quality of wastewater after the treatment process. The quantity and quality of incoming sewage depend on many factors, including weather conditions. Forecasting these parameters can ensure optimal operation of the treatment plant, which will reduce operating costs. For this purpose, using weather data, an attempt was made to estimate the amount of sewage flowing into the sewage treatment plant in Rzeszow. Over 1000 machine learning (ML) models were used, including statistical models such as ARIMA and SARIMAX, and ML algorithms such as KNN and neural networks, in various configurations and time frames. The lowest mean absolute error (MAE) of 3598 m³ and the root mean square error (RMSE) of 4808 m³ were obtained. The study showed how the selection of parameters and different types of predictive models (static, dynamic, machine learning) affects forecast accuracy. It also highlighted that forecasting based solely on basic time-series data is a challenging process.

Keywords: wastewater treatment plants, neural networks, control systems, machine learning

mgr inż. Bartosz Kowal

b.kowal@prz.edu.pl

ORCID: 0000-0002-7909-6484

Asystent w Zakładzie Systemów Złożonych Wydziału Elektrotechniki i Informatyki Politechniki Rzeszowskiej. Jego badania skupiają się na analizie systemów złożonych, bezpieczeństwie komputerowym i sieciowym oraz inżynierii komputerowej.



dr inż. Paweł Kuraś

p.kuras@prz.edu.pl

ORCID: 0000-0002-8658-0821

Adiunkt badawczo-dydaktyczny na Politechnice Rzeszowskiej oraz Akademii Górniczo-Hutniczej, a także wykładowcą Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie, Uniwersytetu WSB Merito w Poznaniu, Wyższej Szkoły Biznesu – National Louis University w Nowym Sączu. Realizował wiele projektów łączących naukę z biznesem, a także jest prezesem pierwszej spółki spin-off Politechniki Rzeszowskiej, Suntrail sp. z o.o.



mgr inż. Patryk Organiściak

org@prz.edu.pl

ORCID: 0000-0002-5277-4038

Jest pracownikiem naukowo-dydaktycznym na Politechnice Rzeszowskiej, gdzie prowadzi zajęcia z technologii informacyjnych, programowania, przetwarzania danych oraz cyberbezpieczeństwa. Jednocześnie realizuje projekty komercyjne w ramach własnego software house, specjalizując się w tworzeniu oprogramowania dla biznesu.



dr hab. inż. Adam Masłoń, prof. PRz

amaslon@prz.edu.pl

ORCID: 0000-0002-3676-0031

Profesor uczelni w Katedrze Inżynierii i Chemii Środowiska Politechniki Rzeszowskiej. Autor i współautor ponad 270 prac naukowych i doniesień konferencyjnych, w tym 45 publikacji indeksowanych w WoS/Scopus. Twórca 15 patentów i pięciu wzorów użytkowych. Recenzent w Narodowym Centrum Badań i Rozwoju oraz Ekspert Regionalnego Programu Operacyjnego Województwa Podkarpackiego. Biegły sądowy z zakresu ochrony środowiska. Członek Regionalnej Komisji ds. Ocen Oddziaływania na Środowisko w Rzeszowie. Wykonawca i kierownik kilkunastu projektów badawczych z technologii ścieków i utylizacji odpadów. Autor ponad 60 opracowań technicznych i ekspertyz dla firm projektowych i przedsiębiorstw komunalnych. Obszar zainteresowań naukowych obejmuje technologie ścieków i utylizację odpadów, efektywność energetyczna w przedsiębiorstwach wodociągowo-kanalizacyjnych, technologie GOZ i bioenergetyczne, w tym technologie biogazowe.



mgr inż. Bartosz Wadiak

bwadiak@mpwik.rzeszow.pl
ORCID: 0009-0004-8894-1952



Absolwent Wydziału Chemicznego Politechniki Rzeszowskiej na kierunku biotechnologia oraz Wydziału Chemii Uniwersytetu Wrocławskiego. Od 2015 r. pracownik Miejskiego Przedsiębiorstwa Wodociągów i Kanalizacji, gdzie początkowo pracował jako analityk w Laboratorium Centralnym, a obecnie zajmuje stanowisko technologa na oczyszczalni ścieków. Odpowiada za realizację i nadzór procesu oczyszczania ścieków komunalnych oraz zagospodarowanie osadów.
