

Bartosz Szelaąg, Lidia Bartkiewicz, Jan Studziński

## Zastosowanie metod czarnej skrzynki do prognozowania wartości wybranych wskaźników jakości ścieków dopływających do oczyszczalni komunalnej

Do modelowania zarówno ilości, jak i jakości ścieków dopływających do oczyszczalni można wykorzystać metody czarnej skrzynki, w których na etapie uczenia modelu tworzona jest jego struktura, decydująca o poprawnym prognozowaniu wartości zmiennej objaśnianej. Autorzy pracy [1] zastosowali z powodzeniem modele typu ARIMA i sztuczne sieci neuronowe (SNN) do symulacji zawartości zawieszin ogólnych oraz azotu i fosforu ogólnego w ściekach dopływających do oczyszczalni. Z kolei w pracy [2] wykazano przydatność modelu ARIMA do prognozowania wartości ChZT i zawartości zawieszin ogólnych w ściekach deszczowych. Jednocześnie opisane w pracy [3] wykorzystanie SNN do prognozowania zawartości związków biogenych w ściekach nie dało zadowalających wyników. Autorzy pracy [4] do prognozowania zawartości zawieszin ogólnych i biochemicznego zapotrzebowania na tlen zastosowali metody wektorów nośnych (SVM), lasów losowych (RF), MARS (multivariate adaptive regression spline) i k-najbliższego sąsiada. Ponadto w pracy tej, ze względu na znaczną liczbę zmiennych wejściowych opisujących wartości wybranych wskaźników jakości ścieków, zastosowano metodę analizy składowych głównych i drzew boostingowych w celu ograniczenia liczby zmiennych w modelu. W praktyce jest to bardzo istotne [5], ponieważ pozwala na wyeliminowanie z obliczeń wartości zmiennych, które nie mają istotnego wpływu na przebieg badanego procesu, a co za tym idzie – upraszcza strukturę tworzonego modelu [6, 7]. Z przeglądu literatury wynika, że w praktyce często stosuje się modele hybrydowe, stanowiące zazwyczaj połączenie modeli czarnej skrzynki (SSN) z odpowiednim modelem klasyfikacyjnym – najczęściej są to samoorganizujące się sieci neuronowe [8, 9]. Obliczenia przeprowadzone przez autorów pracy [7] wykazały, że takie rozwiązanie prowadzi do znacznej poprawy zdolności predykcyjnych modelu. Takie postępowanie ma jednak również wady, szczególnie w przypadku, gdy dysponuje się niewielką liczbą danych wejściowych lub gdy zbiory uzyskane w wyniku klasyfikacji charakteryzuje znaczne zróżnicowanie. W takich przypadkach mogą wystąpić problemy na etapie uczenia oraz testowania modelu regresyjnego.

Dr inż. B Szelaąg: Politechnika Świętokrzyska w Kielcach, Wydział Inżynierii Środowiska, Geomatyki i Energetyki, Katedra Geotechniki, Geomatyki i Gospodarki Odpadami, al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, [bszelaag@tu.kielce.pl](mailto:bszelaag@tu.kielce.pl)

Dr inż. L. Bartkiewicz: Politechnika Świętokrzyska w Kielcach, Wydział Inżynierii Środowiska, Geomatyki i Energetyki, Katedra Technologii Wody i Ścieków, al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce

Dr hab. inż. J. Studziński: Instytut Badań Systemowych Polskiej Akademii Nauk, Centrum Zastosowań Informatyki w Inżynierii Środowiska, ul. Newelska 6, 01-447 Warszawa, [jan.studzinski@ibspan.waw.pl](mailto:jan.studzinski@ibspan.waw.pl)

Z doniesień literaturowych wynika także, że podstawowymi danymi do obliczeń jakości ścieków dopływających do oczyszczalni za pomocą metod czarnej skrzynki są zwykle pięciodobowe biochemiczne zapotrzebowanie na tlen oraz zawartości zawieszin ogólnych i azotu ogólnego. Do prognozowania wartości tych wskaźników wykorzystuje się szeregi czasowe ich wartości uzyskane podczas rutynowych pomiarów fizyczno-chemicznych, tworząc modele o charakterze autoregresyjnym. Jednak ze względu na stosunkowo duże koszty i długi czas pozyskiwania wyników takich pomiarów, celowe jest opracowywanie modeli predykcyjnych na podstawie wskaźników jakości ścieków monitorowanych w oczyszczalni w sposób ciągły i niewymagających stosowania specjalnych technik (podobnie jak pomiar strumienia objętości ścieków). Za takimi modelami przemawia również fakt, że badania opisane w pracach [10, 11] wykazały silną zależność między wartością strumienia objętości ścieków a wartościami wskaźników opisujących ich jakość.

### Obiekt badań

Obiektem badań była miejska oczyszczalnia ścieków w Rzeszowie, wybudowana w latach 1974–1988 i wielokrotnie modernizowana. Projektowana średniodobowa przepustowość tego obiektu wynosiła 62,5 tys. m<sup>3</sup>/d ( $Q_d^{\text{maks}} = 75$  tys. m<sup>3</sup>/d), a równoważna liczba mieszkańców (RLM) była równa 400 tys. Podczas pomiarów prowadzonych w latach 2013–2015 ilość ścieków dopływających do oczyszczalni zmieniała się od 26 973 m<sup>3</sup>/d do 66 773 m<sup>3</sup>/d. Technologia oczyszczania ścieków obejmuje mechaniczno-biologiczne oczyszczanie ścieków ze zintegrowanym usuwaniem substancji biogenych, oparte na konwencjonalnym wielofazowym osadzie czynnym z denitryfikacją wyprzedzającą i nityfikacją w układzie cyrkulacyjnym. Ciąg technologiczny obiektu stanowi krata schodkowa, piaskownik poziomy, osadnik wstępny, komora osadu czynnego oraz osadniki wtórne. Do rozpatrywanej oczyszczalni dopływają ścieki z terenu Rzeszowa, który ma zasadniczo system kanalizacji rozdzielczej, przy czym w części Starego Miasta około 85% kanalizacji stanowi sieć rozdzielcza, a 15%, to sieć ogólnospławna. W niektórych rejonach miasta nie ma kanalizacji deszczowej. Ponad 50% długości kolektorów stanowią kanały betonowe. Około 65% sieci kanalizacyjnej jest eksploatowana przez powyżej 20 lat. Poziom wód gruntowych na terenie miasta silnie zależy od stanów wody w Wisłoku i jego dopływach, które zbierają wody opadowe z Pogórza Dynowskiego. W praktyce, przy

wysokich stanach poziomu wód gruntowych (lokalnie ok. 0,3 m p.p.t.) i dużej ilości wód opadowych, obserwuje się wzrost wypełnień kolektorów.

## Metody

W artykule przedstawiono dwie koncepcje prognozowania wartości wskaźników jakości ścieków dopływających do oczyszczalni. W pierwszym przypadku wyznaczono modele prognostyczne dotyczące wartości biochemicznego zapotrzebowania na tlen (BZT<sub>5</sub>), a także zawartości azotu ogólnego (N<sub>og</sub>) i amonowego (NH<sub>4</sub><sup>+</sup>), fosforu ogólnego (P<sub>og</sub>) oraz zawiesin ogólnych (Z<sub>og</sub>) na podstawie bezpośrednich pomiarów (modele autoregresyjne). W drugim przypadku badano możliwość prognozowania zawartości związków biogenych w ściekach dopływających na podstawie wyników pomiaru ich natężenia przepływu (modele operatorowe). Ten drugi przypadek jest bardzo istotny z praktycznego punktu widzenia, ponieważ takie modele pozwalają znacznie ograniczyć kosztowne badania jakości ścieków. Dodatkowo na uwagę zasługuje fakt, że w przypadku BZT<sub>5</sub> koncepcja modelowania tego wskaźnika jedynie na podstawie mierzonego natężenia przepływu pozwala wyznaczyć jego prognozę w czasie znacznie krótszym niż czas potrzebny do jego oznaczenia (5 d).

Przed przystąpieniem do budowy modeli prognostycznych wejściowe dane pomiarowe poddano modyfikacji polegającej na usunięciu zmiennych mających pomijalny wpływ na przebieg analizowanego zjawiska, przy czym wcześniej dane te poddano normalizacji za pomocą przekształcenia w postaci [12]:

$$\bar{A}_i = \frac{A_i - A_{\min}}{A_{\max} - A_{\min}} \quad (1)$$

w której:

$\bar{A}_i$  – znormalizowana wartość i-tego elementu zbioru A metodą min-maks

$A_i$  – wartość i-tego elementu zbioru A zarejestrowana w czasie pomiarów

$A_{\max}$  – wartość maksymalna w zbiorze A

$A_{\min}$  – wartość minimalna w zbiorze A

Do modyfikacji danych pomiarowych i ograniczenia liczby zmiennych wejściowych modelu zastosowano metodę drzew wzmocnianych (BT) oraz metodę analizy składowych głównych (PCA) [13, 14]. Za pomocą drzew wzmocnianych określono wartości wskaźnika ważności (IMP) poszczególnych predyktorów (zmiennych objaśniających;  $x(t-i)$ ,  $i=1, 2, \dots, 7$ ), przy czym im większa wartość IMP, tym rozpatrywane zmienne mają większy wpływ na zmienną objaśnianą. W metodzie analizy składowych obliczono wartości składowych głównych stanowiących liniową kombinację  $Z = \sum a_i X_i$  analizowanych zmiennych przy założeniu, że otrzymane składowe są ortogonalne. Wyznaczone składowe (a więc również parametry  $a_i$ ) były dobierane w taki sposób, aby uzyskane kombinacje liniowe  $Z_i$  miały największą wariancję wśród wszystkich kombinacji liniowych. Matematyczne rozwiązanie tego problemu otrzymuje się przez znalezienie wartości i wektorów własnych macierzy korelacji. Obliczony wektor  $a_1 = [a_{11}, a_{12}, \dots, a_{1p}]$ , dający maksymalną wariancję przy dodatkowym warunku  $\sum p_i = 1$  ( $a_{1i}^2 = 1$ ), jest wektorem własnym, odpowiadającym największej wartości własnej macierzy korelacji R, opisującej zależności między poszczególnymi wskaźnikami jakości ścieków i natężeniem przepływu ścieków. W dalszych analizach uwzględniono wartości składowych,

w przypadku których wartości własne wynosiły co najmniej 1,0 [16]. W przypadku metody drzew wzmocnianych (BT) wartość graniczną IMP, decydującą o tym, czy badana zmienna będzie ujęta lub pominięta w modelu, wyznaczono analizując zmienność obliczanych wskaźników ważności wybranych predyktorów.

W niniejszym artykule do prognozowania wartości wybranych wskaźników jakości ścieków dopływających do oczyszczalni komunalnej zastosowano metody MARS oraz lasów losowych (RF), a także rozpatrzono wpływ wprowadzenia dodatkowej zmiennej o charakterze klasyfikacyjnym na zdolności predycyjne modelu prognostycznego w metodzie RF.

## Metoda MARS

MARS jest jedną z wielu metod stosowanych do rozwiązywania zagadnień o charakterze regresyjnym [15, 16] i stanowi rozszerzenie klasycznego wykorzystania danych wejściowych przy wyznaczaniu modeli matematycznych. W metodzie tej zakresy zmienności predyktorów branych pod uwagę są dzielone na przedziały, w których rozpatrywane zmienne mogą mieć różny wpływ na badane zjawisko. Granice przedziałów są wyznaczane na podstawie wartości progowych ( $t$ ), co oznacza, że w zależności od tego, czy zmienna jest poniżej czy powyżej wartości  $t$ , rozpatrywany predyktor będzie ujęty w modelu statystycznym z różną wagą lub innym znakiem. Rozróżnianie wartości rozważanych zmiennych objaśniających na mniejsze i większe od wartości progowych ( $t_i$ ) jest realizowane na podstawie funkcji bazowej w postaci:

$$h(X) = \alpha_i (\max(0, X - t)) \quad (2)$$

w której:

$h(X)$  – wektor funkcji bazowych poszczególnych zmiennych ( $x_i$ ), w przypadku których jest spełniony warunek:

$$x_i - t_i = \begin{cases} x_i - t_i, & \text{jeżeli } x_i > t_i \\ 0, & \text{jeżeli } x_i \leq t_i \end{cases} \quad (3)$$

W metodzie tej relację regresyjną stanowi funkcja sklejana, otrzymywana z liniowej kombinacji iloczynu funkcji bazowych z odpowiednimi wagami, którą można zapisać w postaci:

$$f(X) = \alpha_0 + \sum_{m=1}^M \alpha_m h_m(X) \quad (4)$$

w której:

$X = [x_1, x_2, \dots, x_i]$  – wektor danych wejściowych

$\alpha_m$  – wartości wag

$h_m$  – funkcje bazowe

Do wyznaczenia wartości parametrów modelu metodą MARS opracowano specjalny algorytm, który umożliwia przeszukiwanie zbioru obserwacji (danych pomiarowych) w celu określenia wartości progowych (węzłów). Algorytm wykorzystuje metodę rekurencyjnego podziału przestrzeni cech i składa się z występujących naprzemiennie dwóch etapów, dopóki nie zostanie spełnione kryterium zatrzymania (uogólniony błąd w ocenie krzyżowej). W pierwszym etapie algorytmu jest zwiększana złożoność modelu poprzez dodawanie kolejnych funkcji bazowych, dopóki nie zostanie osiągnięta maksymalna liczba funkcji zadeklarowana przez użytkownika. W kolejnym etapie obliczeń jest uruchamiana procedura eliminacji (prycinania), mająca na celu usunięcie z modelu najmniej istotnych funkcji bazowych.

## Metoda lasów losowych (RF)

Algorytm lasów losowych został zaproponowany przez Leo Breimana [17] i jest on rozwinięciem metody bootstrap (szacowanie rozkładu błędów estymacji za pomocą wielokrotnego losowania z powtórzeniami). W pierwszym etapie algorytmu wykonuje się  $k$ -krotne próbkowanie z  $n$ -elementowego zbioru treningowego (dopuszczając powtórzenia), a następnie na podstawie otrzymanych zbiorów tworzone są drzewa regresyjne. Proces ich budowy, w odniesieniu do klasycznego algorytmu lasów losowych, został zmodyfikowany tak, aby w każdym węźle drzewa można było dokonać najlepszego podziału nie na podstawie wszystkich, lecz jedynie wylosowanych atrybutów (zmiennych objaśniających). W ten sposób uzyskuje się  $k$ -drzew regresyjnych tworzących las, na podstawie którego jest wyznaczana prognoza polegająca na obliczeniu średniej arytmetycznej indywidualnych prognoz w przypadku pojedynczych drzew, będąca końcowym wynikiem modelu. Najważniejszymi zaletami lasów losowych są efektywność (wykazana empirycznie [17]) i mała liczba parametrów sterujących ich działaniem, ponieważ użytkownik sam decyduje o tym, z ilu drzew będzie się składał las. Jest to istotna zaleta, w porównaniu do metody sieci neuronowych czy regresyjnej metody wektorów nośnych, w których liczba parametrów (a tym samym liczba możliwych kombinacji ich wartości) jest znaczna.

## Metoda samoorganizujących się sieci neuronowych (SOM)

Najpopularniejszym typem samoorganizujących się sieci neuronowych (SOM – self organization map) są sieci Kohonena [18]. Zalicza się je do sieci uczonych bez nauczyciela, co oznacza, że na etapie uczenia modelu nie są przedstawiane na jego wyjściu żadne wzorce wprowadzonych danych wejściowych. Zadaniem sieci jest utworzenie tych wzorców na etapie uczenia i samodzielne sklasyfikowanie danych wejściowych. Dane treningowe są samodzielnie sklasyfikowane przez sieć na podstawie stopnia ich korelacji. Trenowanie odbywa się więc jako proces samo-uczenia.

Zdolność tego typu sieci neuronowych do odkrywania wzorców sprawia, że są często stosowane do wyszukiwania skupień ukrytych w danych wejściowych, to jest danych charakteryzujących się podobnymi właściwościami. Sieci neuronowe typu SOM składają się z dwóch warstw – wejściowej oraz wyjściowej, określanej również mianem konkurencyjnej, mającej postać jedno- lub dwuwymiarowej tablicy. Sieci te w przeciwieństwie do innych nie mają warstwy ukrytej. W algorytmie stosowanym w większości sieci typu SOM neurony reprezentujące podobne klasy znajdują się obok siebie, tworząc uporządkowaną mapę, dzięki czemu jest możliwe określenie relacji między otrzymanymi klasami. Obliczana sieć neuronowa jest jednokierunkowa, a jej neurony są połączone ze wszystkimi elementami  $N$ -wymiarowego wektora wejściowego ( $x$ ). Wagi połączeń neuronów tworzą wektor  $w_i = [w_{i1}, w_{i2}, \dots, w_{iN}]^T$ . Wektor sygnałów wejściowych przed procesem uczenia jest normalizowany, co można zapisać następująco:

$$\bar{x}_i = \frac{x_i}{\sqrt{\sum_{k=1}^n (x_k)^2}} \quad (5)$$

Po pobudzeniu sieci wektorem wejściowym  $X$  na etapie współzawodnictwa wygra neuron, którego wagi najmniej

różnią się od elementów tego wektora. Zwycięzca,  $w$ -ty neuron, spełnia zależność:

$$d(x, w_m) = \min d(x, w_i) \quad (6)$$

w której:

$w_m$  – wektor wag neuronu zwycięskiego

$d(x, w_i)$  – odległość (najczęściej euklidesowa) między wektorem prezentowanego wzorca wejściowego ( $x$ ) a wektorem wag ( $w_i$ )

$i = 1, 2, \dots, n$  – liczba wyjść

W obrębie  $i$ -tego neuronu przyjmuje się sąsiedztwo  $G(i, x)$ , które w klasycznym algorytmie Kohonena definiowane jest jako  $G(i, x) = \{1, \text{jeżeli } d(i, w) \leq R \text{ i } 0, \text{ jeżeli } d(i, w) > R\}$  ( $R$  – promień sąsiedztwa malejący do 0). W przypadku sieci SOM, wyniki obliczeń stanowi  $n$  zbiorów danych, które zostały poddane klasyfikacji przy wykorzystaniu tego algorytmu. Na etapie uczenia samoorganizujących się sieci neuronowych jest definiowana liczba neuronów tworzących warstwę topologiczną i liczba epok uczenia. W niniejszym artykule samoorganizujące się sieci neuronowe wykorzystano do podziału danych na poszczególne klasy  $i = 1, 2, \dots, n$ . W przypadku, gdy dane wejściowe  $X = [x_1, x_2, \dots, x_i]$  uzyskane metodą PCA lub metodą BT zostały zaliczone do dowolnej klasy, to automatycznie generowano kolejną zmienną ( $u$ ) objaśniającą wartości rozpatrywanych wskaźników jakości ścieków, opisaną jako:

$$u = [u_1, u_2, u_3, \dots, u_n] \quad (7)$$

przy czym  $u_i = 0$  lub  $1$ ,  $i = 1, 2, \dots, n$  oraz jest spełniony warunek  $u_i = 1 \rightarrow u_{n \neq i} = 0$ .

Uzyskana w ten sposób informacja o przydziale do odpowiedniej klasy ( $i$ ) stanowiła dodatkową zmienną ( $u$ ) o charakterze klasyfikacyjnym, objaśniającą wskaźniki jakości ścieków na dopływie do oczyszczalni i ostateczny wektor wejściowy do modelu lasu losowego miał postać  $X' = [X, u]$ . Zaletą takiego podejścia jest fakt, że dane nie są rozdzielane na kolejne zbiory, w których trzeba ponownie dokonywać podziału na zbiór uczący i testowy. W praktyce pozwala to uniknąć problemów na etapie uczenia modelu statystycznego, gdy liczba danych uzyskana na drodze klasyfikacji za pomocą sieci typu SOM jest w jednej z kilku klas stosunkowo niewielka. Na uwagę zasługuje fakt, że przedstawione podejście nie wymaga rozdzielania danych uczących po ich sklasyfikowaniu. Wyznaczony wektor  $u$  stanowi uzupełnienie danych wejściowych zapisanych w postaci wektora  $X$ , dlatego dane wejściowe do modelu (regresyjnego, klasyfikacyjnego) przedstawia wektor w postaci  $X' = [X, u]$ . W artykule badano sieci neuronowe typu Kohonena o warstwie topologicznej  $2 \times 2$ , gdy liczba epok na etapie uczenia wynosiła 1000. W celu prawidłowego przeprowadzenia procesu uczenia, a następnie właściwej oceny działania modelu statystycznego, dokonano podziału danych pomiarowych na zbiór uczący (50%) oraz zbiory testowy i walidacyjny (po 25%).

## Kryteria oceny modeli

Do oceny zdolności predykcyjnej modeli prognozujących wartości wskaźników jakości ścieków dopływających do oczyszczalni (BZT<sub>5</sub>, Z<sub>og</sub>, N<sub>og</sub>, NH<sub>4</sub><sup>+</sup> i P<sub>og</sub>) zastosowano następujące kryteria:

– średni błąd bezwzględny (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{i,obs} - y_{i,progn}| \quad (8)$$

– średni bezwzględny błąd procentowy (MAPE):

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_{i,obs} - y_{i,progn}}{y_{i,obs}} \right| 100 \quad (9)$$

w których:

$n$  – liczebność zbioru danych

$y_{i,obs}, progn$  – odpowiednio: obserwowane (zmierzone) i prognozowane (obliczone) wartości wskaźników jakości ścieków

## Dyskusja wyników

Aby przeprowadzić prawidłowy proces uczenia w analizowanych modelach (metodami RF, RF+SOM i MARS), a następnie właściwą ocenę ich działania, dokonano podziału danych pomiarowych na trzy zbiory, to jest uczyć (50%), walidacyjny (25%) i testowy (25%). W celu ustalenia zakresu stosowności modeli matematycznych w tabeli 1 zestawiono zakresy zmienności wartości analizowanych wskaźników jakości ścieków oraz wartości natężenia ścieków dopływających do oczyszczalni (Q). Na podstawie danych zawartych w tej tabeli można stwierdzić, że zarówno jakość, jak i ilość ścieków trafiających do oczyszczalni w Rzeszowie ulegała znacznym wahaniom, co potwierdził szeroki zakres ich zmienności. Z tego względu wartość obciążenia ładunkiem zanieczyszczeń reaktora biologicznego także podlegała znacznym zmianom. W celu optymalizacji działania reaktora konieczne jest zatem opracowanie modeli dających praktyczną możliwość prognozowania jakości oczyszczanych ścieków.

Tabela 1. Zmienność wartości wskaźników jakości i ilości ścieków dopływających do oczyszczalni komunalnej w Rzeszowie w latach 2013–2015

Table 1. Variability of sewage inflow quality and quantity indicators for the municipal sewage treatment plant in Rzeszow in the period of 2013–2015

Wskaźnik, jednostka	Wartość		
	minimalna	średnia	maksymalna
BZT <sub>5</sub> , gO <sub>2</sub> /m <sup>3</sup>	38,1	374,0	788,0
Zawiesiny ogólne, g/m <sup>3</sup>	80,0	430,0	1140,0
Azot ogólny, gN/m <sup>3</sup>	21,3	69,7	99,0
Azot amonowy, gNH <sub>4</sub> <sup>+</sup> /m <sup>3</sup>	13,7	54,4	80,0
Fosfor ogólny, gP/m <sup>3</sup>	3,4	12,5	37,5
Q, m <sup>3</sup> /d	66973	38658	66773

Tabela 2. Wartości własne macierzy uzyskane metodą analizy składowych głównych (PCA)

Table 2. Eigenvalues of matrices obtained by PCA method

Składowa główna	[BZT <sub>5</sub> ] (BZT <sub>5</sub> )	[BZT <sub>5</sub> ] (Q)	[Z <sub>og</sub> ] (Z <sub>og</sub> )	[Z <sub>og</sub> ] (Q)	[NH <sub>4</sub> <sup>+</sup> ] (NH <sub>4</sub> <sup>+</sup> )	[NH <sub>4</sub> <sup>+</sup> ] (Q)	[N <sub>og</sub> ] (N <sub>og</sub> )	[N <sub>og</sub> ] (Q)	[P <sub>og</sub> ] (P <sub>og</sub> )	[P <sub>og</sub> ] (Q)
PC-1	3,51	4,52	3,44	4,52	4,11	4,52	4,09	4,52	4,04	4,52
PC-2	0,88	0,85	0,86	0,85	0,78	0,85	0,87	0,85	0,77	0,85
PC-3	0,68	0,57	0,64	0,57	0,57	0,57	0,58	0,57	0,54	0,57
PC-4	0,57	0,37	0,62	0,37	0,50	0,37	0,46	0,37	0,49	0,37
PC-5	0,52	0,29	0,55	0,29	0,41	0,29	0,39	0,29	0,46	0,29
PC-6	0,46	0,23	0,46	0,23	0,33	0,23	0,31	0,23	0,36	0,23
PC-7	0,38	0,18	0,43	0,18	0,31	0,18	0,30	0,18	0,35	0,18

Celowe jest także przeprowadzenie oceny możliwości eliminacji zmiennych wejściowych, mających pomijalny wpływ na zmienność analizowanego wskaźnika jakości ścieków. W tym celu, stosując metodę drzew wzmocnianych (BT), wyznaczono wartości wskaźników ważności (IMP) poszczególnych predyktorów oznaczających badane wskaźniki jakości ścieków. Wyniki analiz zamieszczono na rysunku 1, przy czym w lewej kolumnie podano wyniki w przypadku modeli autoregresyjnych, natomiast w prawej wyniki dotyczące danego predyktora w funkcji natężenia ścieków dopływających do oczyszczalni (Q). Przyjęto, że graniczną wartością IMP, decydującą o uwzględnieniu lub nieuwzględnieniu w modelu danej zmiennej wejściowej, jest 0,8. Na podstawie otrzymanych wyników dotyczących poszczególnych zmiennych stwierdzono, że w modelu autoregresyjnym:

– w przypadku BZT<sub>5</sub> istotne były wartości [BZT<sub>5</sub>](t–1) i [BZT<sub>5</sub>](t–3), natomiast w relacji między BZT<sub>5</sub> i Q wartości [Q](t–1), [Q](t–2) oraz [Q](t–3),

– w przypadku zawiesin ogólnych istotne były wartości od [Z<sub>og</sub>](t–1) do [Z<sub>og</sub>](t–4), natomiast w relacji między Z<sub>og</sub> i Q wartości [Q](t–1) i [Q](t–2),

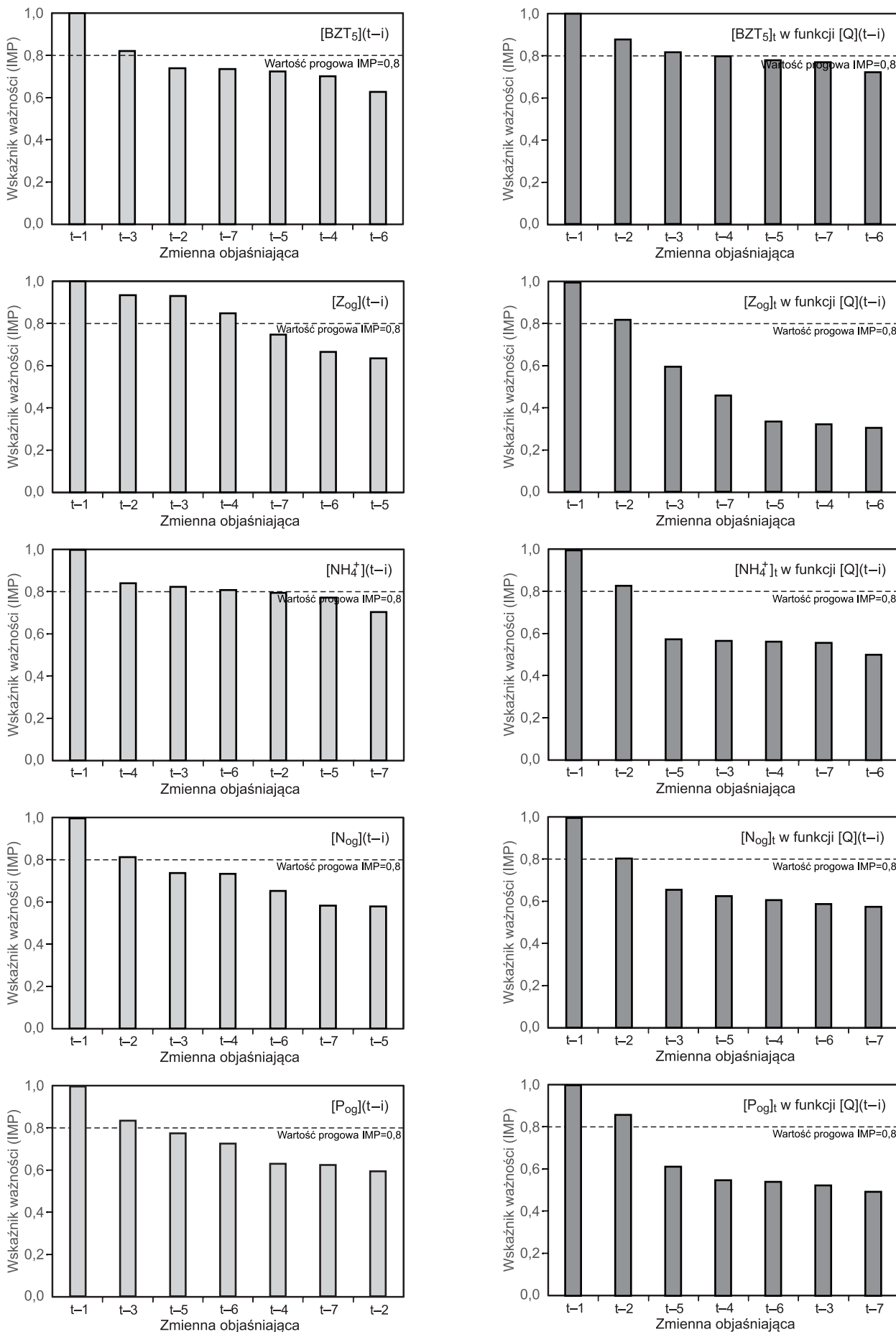
– w przypadku azotu amonowego istotne były wartości od [NH<sub>4</sub><sup>+</sup>](t–1) do [NH<sub>4</sub><sup>+</sup>](t–4), natomiast w relacji między NH<sub>4</sub><sup>+</sup> i Q wartości [Q](t–1) i [Q](t–2), podobnie, jak w przypadku zawiesin ogólnych,

– w przypadku azotu ogólnego istotne były wartości [N<sub>og</sub>](t–1) i [N<sub>og</sub>](t–2), natomiast w relacji N<sub>og</sub> i Q istotne były wartości [Q](t–1) i [Q](t–2),

– w przypadku fosforu ogólnego istotne były wartości [P<sub>og</sub>](t–1) i [P<sub>og</sub>](t–2), natomiast w relacji P<sub>og</sub> i Q wartości [Q](t–1) i [Q](t–2), podobnie, jak w przypadku azotu ogólnego.

Przeprowadzone badania za pomocą metody analizy składowych głównych (PCA) wykazały, że zarówno w przypadku ilości, jak i jakości ścieków na zmienność wartości analizowanych wskaźników jakości ścieków dopływających do oczyszczalni wpływała pierwsza składowa tych zmiennych, na co wskazywały wartości macierzy własnych odpowiednio większe od 1 (tab. 2). Warto zauważyć, że podobny wynik można było otrzymać również w badaniach wykonanych metodą BT, gdyby przyjąć wartość graniczną IMP równą 0,9. Na podstawie wyników analiz istotności zmiennych objaśniających uzyskanych za pomocą metod BT i PCA opracowano modele do prognozowania wartości wskaźników jakości ścieków dopływających do oczyszczalni.

W tabeli 3 zamieszczono zestawienie wartości MAE i MAPE, opisujących dopasowanie wartości prognozowanych do wartości pomierzonych wybranych wskaźników jakości ścieków w przypadku modeli wyznaczonych



Rys. 1. Ranking predyktorów  
Fig. 1. Predictors ranking

Tabela 3. Wartości miar dopasowania (MAE, MAPE) obliczonych (metodami MARS, RF i RF+SOM) wartości wskaźników jakości ścieków do wartości pomierzonych, na podstawie metod BT i PCA  
 Table 3. Accuracy measures (MAE, MAPE) of calculated (by MARS, RF, RF+SOM) sewage quality indicators against the measured values based on BT and PCA methods

Zmienna	MARS				RF				RF+SOM			
	uczenie		test		uczenie		test		uczenie		test	
	MAE, g/m <sup>3</sup>	MAPE, %	MAE, g/m <sup>3</sup>	MAPE, %	MAE, g/m <sup>3</sup>	MAPE, %	MAE, g/m <sup>3</sup>	MAPE, %	MAE, g/m <sup>3</sup>	MAPE, %	MAE, g/m <sup>3</sup>	MAPE, %
metoda drzew wzmacnianych (BT)												
[BZT <sub>5</sub> ](BZT <sub>5</sub> )	42,25	11,89	45,22	13,53	41,15	11,02	43,78	13,26	29,81	9,21	32,36	10,09
[BZT <sub>5</sub> ](Q)	44,12	14,01	48,32	14,87	42,71	13,49	46,78	14,32	40,97	13,30	44,87	14,12
[Z <sub>og</sub> ](Z <sub>og</sub> )	57,12	16,31	63,24	17,06	59,24	16,90	65,59	17,68	44,72	12,55	49,51	13,13
[Z <sub>og</sub> ](Q)	66,15	18,21	71,99	20,10	71,32	19,27	77,62	21,27	62,25	18,38	68,71	20,00
[N <sub>og</sub> ](N <sub>og</sub> )	6,00	9,71	6,59	10,44	5,95	9,56	6,54	10,28	4,68	7,78	5,14	8,36
[N <sub>og</sub> ](Q)	6,42	10,37	7,05	11,15	6,25	10,03	6,87	10,78	6,01	9,58	6,60	10,30
[NH <sub>4</sub> <sup>+</sup> ](NH <sub>4</sub> <sup>+</sup> )	5,00	11,05	5,84	12,29	9,16	10,70	5,66	11,90	4,14	9,30	4,83	10,34
[NH <sub>4</sub> <sup>+</sup> ](Q)	5,08	11,03	5,93	12,27	5,00	10,75	5,84	11,96	4,68	10,70	5,47	11,90
[P <sub>og</sub> ](P <sub>og</sub> )	1,52	13,25	1,70	14,89	1,60	12,36	1,65	13,69	1,26	11,07	1,41	12,44
[P <sub>og</sub> ](Q)	1,92	16,91	2,15	19,00	1,89	16,54	2,11	18,86	1,87	16,78	2,09	18,59
metoda analizy składowych głównych (PCA)												
[BZT <sub>5</sub> ](BZT <sub>5</sub> )	43,15	12,02	46,25	14,15	38,20	12,84	44,97	13,76	19,07	6,06	20,44	6,50
[BZT <sub>5</sub> ](Q)	44,12	14,01	52,91	16,57	42,97	12,98	50,82	15,57	17,34	5,71	20,80	6,75
[Z <sub>og</sub> ](Z <sub>og</sub> )	57,12	16,31	64,57	17,78	55,84	16,18	63,12	17,64	27,74	7,65	31,36	8,34
[Z <sub>og</sub> ](Q)	66,15	18,21	75,74	21,20	69,50	19,33	72,58	22,50	28,96	7,56	33,72	8,66
[N <sub>og</sub> ](N <sub>og</sub> )	6,00	9,71	6,88	11,12	5,84	9,57	6,70	10,96	2,73	4,37	3,13	5,00
[N <sub>og</sub> ](Q)	7,04	11,40	8,07	13,06	6,84	11,18	7,84	12,80	3,03	4,99	3,48	5,71
[NH <sub>4</sub> <sup>+</sup> ](NH <sub>4</sub> <sup>+</sup> )	5,00	11,05	5,96	12,73	4,85	10,89	5,78	12,54	2,27	4,97	2,70	5,72
[NH <sub>4</sub> <sup>+</sup> ](Q)	5,97	13,19	7,12	15,20	5,75	12,79	7,85	14,74	2,33	5,16	2,78	5,95
[P <sub>og</sub> ](P <sub>og</sub> )	1,52	13,25	1,79	16,02	1,50	12,90	1,77	15,60	0,98	8,23	1,15	9,95
[P <sub>og</sub> ](Q)	2,00	17,19	2,36	20,78	1,90	16,83	2,24	20,35	1,10	9,42	1,29	11,39

metodami MARS, RF (lasów losowych) oraz RF+SOM (lasów losowych i sieci neuronowych). Na podstawie wartości błędów w przypadku modeli otrzymanych metodą drzew wzmacnianych (BT) można stwierdzić, że nieznacznie lepsze wyniki obliczeń uzyskano za pomocą modeli autoregresyjnych, w których wartość danego wskaźnika wyznacza się na podstawie jego wcześniejszych wartości. Przykładowo, w przypadku prognozy wartości BZT<sub>5</sub>, wyznaczonej metodą MARS na podstawie [BZT<sub>5</sub>](t-1) i [BZT<sub>5</sub>](t-2), błędy prognozy tego wskaźnika MAE i MAPE wyniosły odpowiednio 45,22 gO<sub>2</sub>/m<sup>3</sup> i 13,53%, natomiast błędy obliczone na podstawie wartości natężenia ścieków dopływających do oczyszczalni [Q](t-1) i [Q](t-2) były równe odpowiednio 48,32 gO<sub>2</sub>/m<sup>3</sup> i 14,87%. Wśród rozpatrywanych trzech metod modelowania najlepsze wyniki prognozy jakości ścieków otrzymano metodą lasów losowych (RF), która uwzględnia dodatkową zmienną klasyfikacyjną wyznaczoną za pomocą sieci typu SOM, co dokumentują wartości błędów dopasowania wyników modelowania do wyników pomiarów (w przypadku modeli [N<sub>og</sub>](Q), metodą RF+SOM otrzymano MAE=6,60 gN/m<sup>3</sup> i MAPE=10,30%, metodą MARS – MAE=7,05 gN/m<sup>3</sup> i MAPE=11,15%, zaś metodą RF – MAE=6,87 gN/m<sup>3</sup> i MAPE=10,78%). Uwzględnienie dodatkowej zmiennej o charakterze klasyfikacyjnym w modelach lasów losowych

(RF+SOM) uzyskanych na podstawie [Q](t-i) spowodowało zmniejszenie wartości MAE w przypadku BZT<sub>5</sub> o 4,1% (z 46,78 gO<sub>2</sub>/m<sup>3</sup> do 44,87 gO<sub>2</sub>/m<sup>3</sup>), w przypadku zawiesin ogólnych o 13% (z 77,62 g/m<sup>3</sup> do 68,71 g/m<sup>3</sup>), w przypadku azotu ogólnego o 4,0% (z 6,87 gN/m<sup>3</sup> do 6,60 gN/m<sup>3</sup>), w przypadku azotu amonowego o 6,8% (z 5,84 gNH<sub>4</sub><sup>+</sup>/m<sup>3</sup> do 5,47 gNH<sub>4</sub><sup>+</sup>/m<sup>3</sup>) oraz w przypadku fosforu ogólnego o 1,0% (z 2,11 gP/m<sup>3</sup> do 2,09 gP/m<sup>3</sup>). Odpowiednie zmniejszenia wartości MAPE wynosiły 1,4% (BZT<sub>5</sub>), 6,0% (Z<sub>og</sub>), 4,5% (N<sub>og</sub>), 0,5% (NH<sub>4</sub><sup>+</sup>) oraz 1,4% (P<sub>og</sub>). W przypadku modeli opracowanych metodami MARS i RF wartości MAE i MAPE różniły się tylko nieznacznie. Przykładowo, podczas autoregresyjnego modelowania ilości zawiesin ogólnych w ściekach na podstawie [Z<sub>og</sub>](t-1) i [Z<sub>og</sub>](t-2), błędy prognozy w przypadku metody MARS wynosiły MAE=63,24 g/m<sup>3</sup> i MAPE=17,06%, zaś w przypadku metody RF – MAE=65,59 g/m<sup>3</sup> i MAPE=17,68% (dane testowe).

Nieznaczne różnice w wartościach MAE i MAPE otrzymano także w przypadku modeli wyznaczonych metodami MARS i RF, na podstawie wyników ustalonych z użyciem metody analizy składowych głównych (PCA). Podobnie jak poprzednio, obliczone wartości błędów prognozy różniły się nieznacznie, natomiast wprowadzenie dodatkowej zmiennej klasyfikującej dane wejściowe, ustalonej za

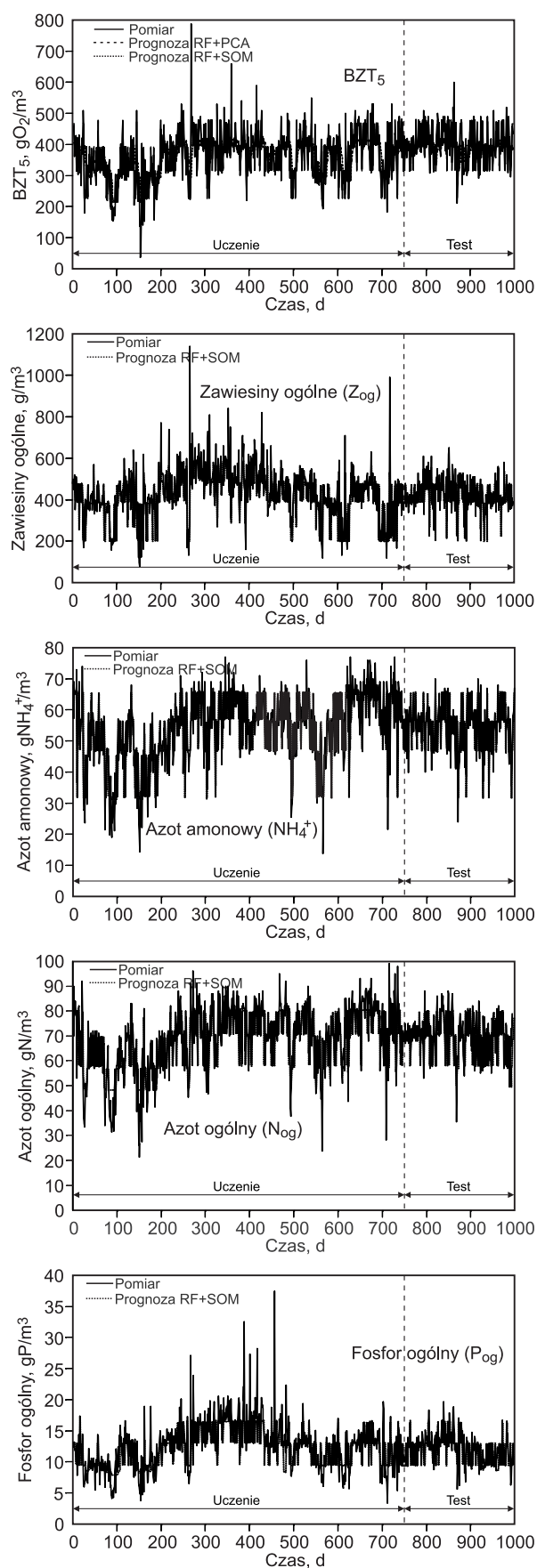
pomocą sieci typu SOM, miało istotny dodatni wpływ na wyniki prognozy wartości wskaźników jakości ścieków, przy czym był on znacznie większy, niż w poprzedniej analizie. Na przykład, zastosowanie metody RF+SOM wpłynęło na zmniejszenie wartości MAE w modelach autoregresyjnych, w porównaniu z metodą RF, odpowiednio o 54,6% w przypadku BZT<sub>5</sub> (z 44,97 gO<sub>2</sub>/m<sup>3</sup> do 20,44 gO<sub>2</sub>/m<sup>3</sup>), o 50,3% w przypadku zawiesin ogólnych (z 63,12 g/m<sup>3</sup> do 31,36 g/m<sup>3</sup>), o 53,3% w przypadku azotu ogólnego (z 6,70 gN/m<sup>3</sup> do 3,13 gN/m<sup>3</sup>), o 53,3% w przypadku azotu amonowego (z 5,78 gNH<sub>4</sub><sup>+</sup>/m<sup>3</sup> do 2,70 gNH<sub>4</sub><sup>+</sup>/m<sup>3</sup>) oraz o 35,0% w przypadku fosforu ogólnego (z 1,77 gP/m<sup>3</sup> do 1,15 gP/m<sup>3</sup>).

W celu wizualizacji wyników modelowania na rysunku 2 zamieszczono przykładowe porównanie wyników pomiarów i prognoz otrzymanych za pomocą metod RF i RF+SOM (BZT<sub>5</sub>) oraz metody RF+SOM (Z<sub>og</sub>, NH<sub>4</sub><sup>+</sup>, N<sub>og</sub> i P<sub>og</sub>), przy czym zmienne objaśniające ustalono na podstawie metody analizy składowych głównych (PCA). Porównując wyniki obliczeń zamieszczone w tabeli 3 można założyć, że zastosowanie analizy PCA w przypadku modeli opracowanych za pomocą metody MARS prowadziło do nieznacznego pogorszenia wyników modelowania w porównaniu z modelami, w których predyktory ustalono metodą drzew wzmacnianych (BT). Przykładowo, w przypadku modelu dotyczącego fosforu ogólnego, wyznaczonego na podstawie wartości natężenia przepływu ścieków [Q](t-1) i [Q](t-2), ustalonych za pomocą metody BT, wartości błędów prognozy (MAE i MAPE) wyniosły 2,15 gP/m<sup>3</sup> i 19,00%, natomiast w przypadku modelu utworzonego z uwzględnieniem składowej PC-1, ustalonej za pomocą metody PCA, wyniosły odpowiednio 2,36 gP/m<sup>3</sup> i 20,78%.

W modelach opisujących wartości takich wskaźników jakości ścieków, jak BZT<sub>5</sub>, azot ogólny, azot amonowy oraz fosfor ogólny, uzyskanych za pomocą metody lasów losowych (RF), gdy predyktory ustalono za pomocą metody drzew wzmacnianych (BT), błędy prognozy były nieznacznie mniejsze, niż gdy zmienne wejściowe ustalono metodą analizy składowych głównych (PCA). W przypadku modelu opisującego zawartość zawiesin ogólnych w ściekach wystąpiła sytuacja odwrotna, to znaczy dobór zmiennych wejściowych na podstawie metody PCA nieznacznie wpłynął na poprawę zdolności prognostycznych modelu (tab. 3). Z kolei w modelach wyznaczanych za pomocą metody RF+SOM zastosowanie metody PCA zamiast metody BT do ograniczenia liczby zmiennych wejściowych prowadziło zawsze do zmniejszenia wartości błędów prognozy modelowanych wartości wskaźników jakości ścieków. Przykładowo, w przypadku modelu opisującego zawartość azotu amonowego w ściekach, uwzględniającego predyktory [Q](t-i) ustalone metodą BT, wartości błędów prognozy wyniosły MAE=5,47 gNH<sub>4</sub><sup>+</sup>/m<sup>3</sup> i MAPE=11,90%; natomiast w przypadku, gdy w modelu tym uwzględniono predyktory [Q](t-i) ustalone metodą PCA, wartości błędów były już prawie dwukrotnie mniejsze i wyniosły MAE=2,78 gNH<sub>4</sub><sup>+</sup>/m<sup>3</sup> i MAPE=5,95%.

## Podsumowanie

Na podstawie obliczeń symulacyjnych wykazano, że korzystając jedynie z pomiarów natężenia ścieków dopływających do oczyszczalni komunalnej można modelować i prognozować zmiany wartości wybranych wskaźników jakości oczyszczanych ścieków. W przeprowadzonych obliczeniach zastosowano metody MARS i lasów losowych



Rys. 2. Porównanie wyników pomiarów i obliczeń wartości wybranych wskaźników jakości ścieków metodami RF oraz RF+SOM (zmienne objaśniające ustalono metodą PCA)  
Fig. 2. Comparison of the measurement data and calculations of selected sewage quality indicator values obtained by RF and RF+SOM methods (independent variables determined by PCA)

(RF) oraz metodę RF+SOM, będącą połączeniem metody lasów losowych (RF) z metodą sieci neuronowych (SOM). Badanymi i porównywanymi modelami opisującymi jakość ścieków (biochemiczne zapotrzebowanie na tlen, zawiesiny ogólne, azot amonowy i azot ogólny oraz fosfor ogólny) były modele autoregresyjne, w których prognozowany parametr zależy od wartości tego parametru w chwilach wcześniejszych, a także modele operatorowe, będące relacją między prognozowanym składnikiem zanieczyszczeń oraz zmierzonymi odpowiednio wcześniej wartościami natężenia ścieków dopływających do oczyszczalni. W przypadku obu rodzajów modeli istotny był dobór predyktorów, czyli odpowiednio przesuniętych w czasie zmiennych wejściowych. W tym celu zastosowano dwie metody doboru zmiennych wejściowych – metodę drzew wzmacnianych oraz (BT) metodę analizy składowych głównych (PCA).

Z przeprowadzonych badań wynika, że modele prognostyczne wartości wybranych wskaźników jakości ścieków wyznaczane na podstawie wartości natężenia ścieków dopływających do oczyszczalni (modele operatorowe) były na ogół niewiele gorsze od modeli autoregresyjnych. Metody MARS i RF dały zbliżone wyniki, jednak znacznie gorsze od metody mieszanej RF+SOM. Modele, w których zmienne wejściowe ustala się za pomocą metody PCA są zwykle lepsze, niż gdy dobór predyktorów następuje za pomocą metody BT. W przypadku, gdy do prognozowania stosowano metodę RF+SOM, a zmienne wyjściowe były dobierane za pomocą metody PCA, praktycznie nie było różnic jakościowych między wynikami modeli autoregresyjnych i modeli operatorowych – najlepsze wyniki otrzymano prognozując wartości BZT<sub>5</sub> i azotu amonowego (różnice w dokładności tych modeli wynosiły 2÷3%, a w przypadku modeli dotyczących pozostałych wskaźników jakości ścieków nie przekraczały 12%).

## LITERATURA

1. S.A. DELLANA, D. WEST: Predictive modeling for wastewater applications: Linear and nonlinear approaches. *Environmental Modelling and Software* 2009, Vol. 24, pp. 96–106.
2. A.G. EI-DIN, D.W. SMITH: Modelling approach for high flow rate in wastewater treatment operation. *Journal of Environmental Engineering and Science* 2002, Vol. 1, No. 4, pp. 275–291.
3. D. RAHA: Exploring artificial neural networks (ANN) modelling for a biological nutrient removal (BNR) sewage treatment plant (STP) to forecast effluent suspended solids. *Indian Chemical Engineer* 2007, Vol. 49, No. 3, pp. 205–220.
4. A. VERMA, X. WEI, A. KUSIAK: Predicting the total suspended solids in wastewater: A data-mining approach. *Engineering Applications of Artificial Intelligence* 2013, Vol. 26, No. 4, pp. 1366–1372.
5. K.P. OLIVEIRA-ESQUERRE, M. MORI, R.E. BRUNS: Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis. *Brazilian Journal of Chemical Engineering* 2002, Vol. 19, No. 4, pp. 365–370.
6. L. RUTKOWSKI: Metody i techniki sztucznej inteligencji. PWN, Warszawa 2006.
7. M. CHUCHRO: Struktura szeregu czasowego natężenia dopływu ścieków do oczyszczalni. Rozprawa doktorska, Akademia Górniczo-Hutnicza, Kraków 2009.
8. H.G. HAN, J.F. QIAO: Prediction of activated sludge bulking based on a self-organizing RBF neural network. *Journal of Process Control* 2012, Vol. 22, No. 6, pp. 1103–1112.
9. H.G. HAN, Y. LI, Y. GUO, J.F. QIAO: A soft computing method to predict sludge volume index based on a recurrent self-organizing neural network. *Applied Soft Computing* 2016, Vol. 38, pp. 477–486.
10. G. KACZOR, T. BERGEL, P. BUGAJSKI: Impact of extraneous waters on the proportion of sewage pollution indices regarding its biological treatment. *Infrastruktura i Ekologia Terenów Wiejskich* 2015, nr IV/3, ss. 1251–1260.
11. G. KACZOR: Oddziaływanie wód przypadkowych na stężenie związków biogenych w ściekach surowych i oczyszczonych podczas pogody mokrej. *Infrastruktura i Ekologia Terenów Wiejskich* 2012, nr 3/IV, ss. 179–191.
12. J. STUDZIŃSKI, L. BARTKIEWICZ, M. STACHURA: Development of mathematical models for forecasting hydraulic loads of water and wastewater networks. *EnviroInfo 2013: Environmental Informatics and Renewable Energies*, Shaker Verlag, Aachen 2013, pp. 1–13.
13. I. T. JOLLIFFE: Principal Component Analysis. Springer Series in Statistics, 2<sup>nd</sup> edition, Springer, New York 2002.
14. T. KUDO, Y. MATSUMOTO: A boosting algorithm for classification of semistructured text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2004 (EMNLP)*, pp. 301–308.
15. W. ZHANG, A. T. C. GOH: Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers* 2016, Vol. 7, No. 1, pp. 45–52.
16. G. GUTIÉRREZ, Á.S. SCHANBEL, J.F.L. CONTADOR: Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecological Modelling* 2009, Vol. 220, No. 24, pp. 3630–3637.
17. L. BREIMAN: Random forest. *Journal Machine Learning* 2000, Vol. 45, No. 1, pp. 5–32.
18. T. KOHONEN: Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 1982, Vol. 43, pp. 59–69.

**Szeląg, B., Bartkiewicz, L., Studziński, J. Black-box Forecasting of Selected Indicator Values for Influent Wastewater Quality in Municipal Treatment Plant. *Ochrona Środowiska* 2016, Vol. 38, No. 4, pp. 39–46.**

**Abstract:** Forecasting the amount and quality of wastewater flowing into a treatment plant sufficiently in advance, enables effective control of numerous treatment process parameters. Therefore, mathematical (physical deterministic and time series statistical) models forecasting both the amount and quality of wastewater inflow into a sewage treatment plant are under development. In this paper, a possibility of simpler time series models application to forecasting values of selected indicators (biochemical oxygen demand (BOD<sub>5</sub>), total suspended solids (TSS), total nitrogen (TN), total phosphorus (TP) and ammonium (NH<sub>4</sub><sup>+</sup>)) of sewage quality in the inflow into a treatment plant

was investigated. The research was based solely on sewage flow rate data and – for the purpose of comparison – the actual measured indicator values. For this purpose, MARS type black-box and random forest (RF) methods were used. Also, a possibility of combining the RF method with a classification model (RF+SOM) was investigated. Boosted trees (BT) and principal component analysis (PCA) methods were applied for identification of data that determine variability of the selected sewage quality indicators. The models were developed on the basis of continuous daily measurements performed in the period of 2013–2015 in the municipal sewage treatment plant in Rzeszów.

**Keywords:** Sewage, modeling, sewage quality forecasting, MARS (multivariate adaptive regression spline), random forest (RF), self-organizing map (SOM), boosted trees (BT), principal component analysis (PCA).