drzewa regresyjne, przewód wodociągowy,
intensywność uszkodzeń, przewidywanie

Małgorzata KUTYŁOWSKA*

DRZEWA REGRESYJNE JAKO NARZĘDZIE DO PRZEWIDYWANIA AWARYJNOŚCI PRZEWODÓW WODOCIĄGOWYCH

Praca przedstawia wyniki przewidywania, za pomocą metody drzew regresyjnych, wskaźnika intensywności uszkodzeń przewodów magistralnych, rozdzielczych i przyłączy wodociągowych w wybranym mieście Polski. Podczas modelowania zbudowano kilka modeli drzew regresyjnych. Wyboru modeli optymalnych (oddzielnie dla każdego typu przewodu wodociągowego) dokonano na zasadzie analizy tzw. kosztów. Struktura drzewa regresyjnego zawierała zmienne niezależne, tzw. predyktory (liczba uszkodzeń i długość przewodów wodociągowych). Zmienną zależną były wskaźniki awaryjności trzech typów przewodów. Modele optymalne charakteryzowały się najmniejszymi kosztami oraz relatywnie prostą architekturą drzewa. Dane eksploatacyjne z lat 2005–2012 posłużyły do wyznaczenia rzeczywistych wartości wskaźnika intensywności uszkodzeń, a także do budowy modeli drzew regresyjnych. Modele optymalne do przewidywania awaryjności przewodów rozdzielczych i przyłączy zawierały 3 węzły dzielone i 4 końcowe, natomiast drzewo regresyjne do modelowania awaryjności przewodów magistralnych było mniej złożone, zawierało 1 węzeł dzielony i 2 końcowe. Uzyskane zbieżności danych rzeczywistych z przewidywanymi można uważać, z inżynierskiego punktu widzenia, za satysfakcjonujące.

1. WPROWADZENIE

Awaryjność sieci wodociągowych jest jednym z kilku wskaźników branych pod uwagę podczas oceny niezawodności działania systemów dystrybucji wody [3, 7, 9]. Obecnie coraz większą wagę przykłada się nie tylko do wyznaczenia wskaźnika intensywności uszkodzeń przewodów wodociągowych na podstawie jedynie danych eksploatacyjnych, ale również do możliwości jego przewidywania za pomocą dostępnych modeli i narzędzi matematycznych [1, 8]. Modelowanie matematyczne oczywiście musi

* Katedra Wodociągów i Kanalizacji, Wydział Inżynierii Środowiska, Politechnika Wrocławska, Wyb. Wyspiańskiego 27, 50–370 Wrocław, malgorzata.kutylowska@pwr.edu.pl.

być poprzedzone zgromadzeniem danych eksploatacyjnych, które stanowią bazę informacji o rozpatrywanej sieci wodociągowej. Istnieje wiele typowych rozwiązań i modeli matematycznych [6, 11, 12], które pozwalają na prognozowanie awaryjności rurociągów, a tym samym ułatwiają ocenę niezawodności ich działania i wpływają na możliwość szybkiej reakcji w chwilach wystąpienia uszkodzenia.

Ostatnio coraz częściej stosuje się wiele metod regresyjnych do rozwiązywania problemów inżynierskich. Do algorytmów regresyjnych należą między innymi: metoda wektorów nośnych (SVM), za pomocą której oszacowano lokalizację wycieków z sieci wodociągowej [2], metoda K-najbliższych sąsiadów (KNN), którą dokonano analizy szeregów czasowych w szeroko pojętych procesach przemysłowych [4] oraz metoda drzew regresyjnych (RT) zastosowana w dziedzinie ekonomii [5].

Drzewa regresyjne i klasyfikacyjne są wykorzystywane odpowiednio do przewidywania zmiennych ilościowych i jakościowych. Początek stosowania tej metody analizy i przewidywania danych datuje się na lata 60. XX wieku, jednak dopiero w 1984 roku L. Breiman spopularyzował tę dziedzinę w książce „Classification and Regression Trees". Ogólnie rzecz ujmując, drzewo regresyjne lub klasyfikacyjne (RT) jest grafem skierowanym, zawierającym korzeń i węzły (liście), w których sprawdzane są warunki dotyczące zmiennych, a także gałęzie zawierające reguły decyzyjne. Analiza wykorzystująca algorytm budowy drzew polega na znalezieniu zbioru logicznych warunków podziału. Zaletą stosowania drzew jest relatywnie prosta interpretacja wyników oraz dobre rezultaty predykcji [13].

Celem niniejszej pracy jest ukazanie możliwości zastosowania RT do przewidywania wskaźnika intensywności uszkodzeń rurociągów magistralnych, rozdzielczych i przyłączy domowych przykładowo wybranej sieci wodociągowej. Do tej pory metodyka ta nie była szeroko stosowana w analizie awaryjności i niezawodności działania systemów komunalnych w Polsce, co stało się przyczynkiem do podjęcia tego tematu.

2. METODYKA BADAŃ

Przewidywanie wskaźnika intensywności uszkodzeń (λ, uszk./(km·a)) przewodów wodociągowych przeprowadzono z wykorzystaniem metody drzew regresyjnych. Dokonano oddzielnie predykcji wskaźnika awaryjności przewodów magistralnych (λ_m), rozdzielczych (λ_r) i przyłączy (λ_p), co oznacza konieczność budowy trzech różnych modeli drzew. Wskaźniki λ były zmiennymi zależnymi, natomiast predyktorami (zmiennymi niezależnymi) długość (L_m, L_r, L_p) i liczba uszkodzeń (N_m, N_r, N_p), odpowiednio przewodów magistralnych, rozdzielczych i przyłączy. Dane eksploatacyjne, uzyskane z przedsiębiorstwa wodociągowego z lat 2005–2012 [10], posłużyły do wyznaczenia rzeczywistego wskaźnika λ oraz do przewidywania wskaźnika awaryjności za pomocą

metody drzew regresyjnych. Niniejsza praca ma charakter wstępny i jej celem jest wskazanie ogólnych możliwości stosowania metody RT. W związku z tym stworzone modele zawierały podstawowe dane (zmienne niezależne L i N), aby ocenić, czy proponowane podejście jest właściwe. Kolejnym etapem analiz będzie zwiększenie wektora zmiennych niezależnych. Jednak na obecnym etapie badań celem tej pracy było wykazanie (na relatywnie prostym przykładzie), czy metoda drzew regresyjnych w ogóle może być stosowana do przewidywania wskaźnika awaryjności przewodów wodociągowych. Należy również pamiętać, że drzewa zbyt złożone są trudne w interpretacji, więc tak samo jak w każdym innym podejściu, tak i w przypadku metody RT, dąży się do budowy modeli jak najprostszych. Zakres zmian zmiennych zależnych i predyktorów przedstawiono w tabeli 1.

Tabela 1. Zakres zmian wartości zmiennych zależnych i predyktorów w latach 2005–2012

Zmienna	L_m, km	L_r, km	L_p, km	N_m, szt.	N_r, szt.	N_p, szt.
MIN	201,7	1197,3	401,9	16	328	109
MAX	212,4	1266,5	434,4	35	518	380
Zmienna	λ_m, uszk./(km·a)	λ_r, uszk./(km·a)	λ_p, uszk./(km·a)			
MIN	0,08	0,26	0,26			
MAX	0,16	0,43	0,93			

Wybór optymalnego modelu drzewa regresyjnego został dokonany na podstawie tzw. resubstytucji kosztów, gdzie obliczany jest oczekiwany błąd kwadratowy wg zależności [13]:

$$R(d) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - d(x_i)\right)^2 \tag{1}$$

w której próba ucząca Z składa się z punktów (x_i, y_i), dla $i = 1, 2,..., N$. Obliczenia przeprowadza się dla tego samego zbioru danych, na podstawie którego zbudowano model d [13]. Pojęcie tzw. kosztu (w metodzie RT [13]) jest uogólnieniem idei, że najlepszą predykcją charakteryzuje się model o najmniejszym błędzie. Miarą kosztu jest stosunek błędnie zdefiniowanych przypadków do wszystkich przypadków. Zatem model optymalny powinien charakteryzować się najmniejszym kosztem. Struktura drzewa (liczba gałęzi i węzłów) zależy od liczby podziałów, która będzie odpowiadała za najlepszą predykcję. Podziały są dokonywane do momentu, gdy węzły są jednorodne lub zawierają określoną liczbę przypadków. Ważnym elementem analizy i doboru wielkości drzewa jest wykonanie tzw. V-krotnego sprawdzianu krzyżowego, który polega na losowym podziale danych uczących. Wymagane jest wielokrotne tworzenie danego

drzewa, a także duży zbiór danych, umożliwiający wykonanie wspomnianych wyżej podziałów. Obliczenia zaprezentowane w niniejszej pracy przeprowadzono w programie Statistica 12.0.

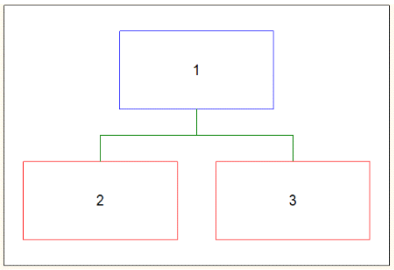
3. WYNIKI I DYSKUSJA

Rzeczywisty wskaźnik intensywności uszkodzeń przewodów magistralnych, rozdzielczych i przyłączy, w latach 2005–2012, wynosił średnio 0,13; 0,35 i 0,69 uszk./(km·a). W oparciu o metodykę przedstawioną powyżej, wybrano optymalne modele drzew regresyjnych, których struktury przedstawiono na rysunku 1. W przypadku modelu drzewa do przewidywania awaryjności przewodów magistralnych liczba węzłów dzielonych była równa 1, a końcowych 2. Natomiast struktury drzew optymalnych do modelowania awaryjności przewodów rozdzielczych i przyłączy były takie same i posiadały liczbę węzłów dzielonych równą 3, a końcowych 4. Podział na kolejne gałęzie uzależniony był od wartości liczby uszkodzeń w przypadku analizy przewodów magistralnych i przyłączy. Dla przewodów magistralnych tą graniczną wartością dla węzła 1 była liczba 24,5, natomiast dla przyłączy liczby 186,5; 360,0 oraz 275,0, odpowiednio dla węzłów numer 1, 3 oraz 4. W odniesieniu do przewodów rozdzielczych liczba uszkodzeń oraz długość były zmiennymi, które brały udział w podziale i wynosiły odpowiednio 355,0 (węzeł 1) i 491,0 (węzeł 3) oraz 1235,9 km (węzeł 4).

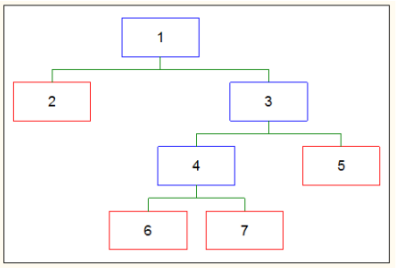
W tabeli 2 zestawiono minimalne koszty dla optymalnych modeli. Pozostałe struktury drzew regresyjnych charakteryzowały się kosztami większymi o rząd wielkości (lub nawet dwa). Przykładowy wykres sekwencji kosztów, dla kilku zbudowanych modeli drzew regresyjnych, przedstawiono na rysunku 2. Analogiczna analiza minimalnych kosztów została przeprowadzona w odniesieniu do pozostałych dwóch typów przewodów wodociągowych. Analiza rysunku 2 wskazuje, że drzewo numer 1 charakteryzowało się najmniejszym kosztem (0,000435), a zatem to właśnie ten model był optymalny do przewidywania wskaźnika awaryjności przyłączy.

Dla każdego węzła została obliczona średnia wartość wskaźnika awaryjności oraz wariancja. Wartości te wraz z licznością próby przypadającej na każdy węzeł przedstawiono w tabeli 3. Liczność ta określa, ile przypadków było analizowanych w danym węźle. W pracy przedstawiono prosty przykład zastosowania drzew regresyjnych w zagadnieniach przewidywania zmiennych opisujących stan techniczny przewodów wodociągowych.

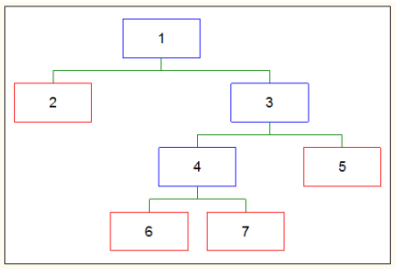
a)



b)



c)

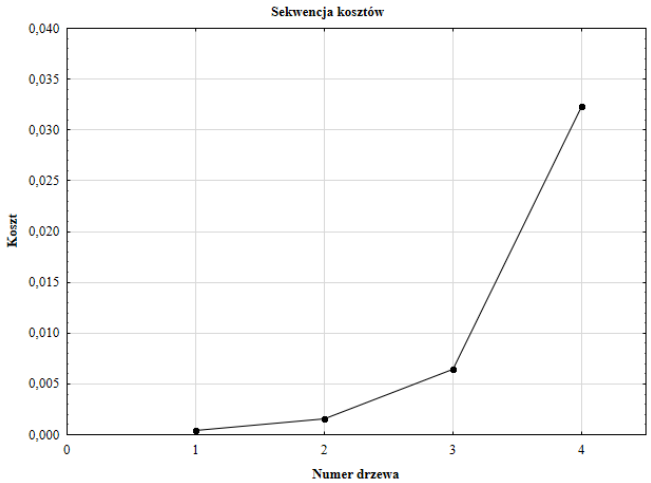


Rys. 1. Optymalne struktury drzew regresyjnych: a) przewody magistralne, b) przewody rozdzielcze, c) przyłącza

Tabela 2. Resubstytucja kosztów dla modelu optymalnego

Rodzaj przewodów	Koszt
Magistralne	0,000118
Rozdzielcze	0,000058
Przyłącza	0,000435

Należy wspomnieć, że drzewa regresyjne mogą być łączone w całe zespoły drzew, tworząc tzw. losowy las. Zespół drzew daje z reguły lepsze wyniki przewidywania niż jedno, nawet najbardziej skomplikowane drzewo [13]. W takim przypadku konieczne byłoby włączenie do wektora predyktorów wielu innych zmiennych, aby skomplikowana architektura modelu miała przełożenie na relacje między zmienną przewidywaną a zmiennymi niezależnymi. Jest to oczywiście limitowane możliwością uzyskania wielu zmiennych eksploatacyjnych. W przyszłości właściwe wydaje się uzupełnienie wektora predyktorów o informacje dotyczące ciśnienia panującego w sieci wodociągowej oraz o inne dane dotyczące rurociągów (np. materiał, średnica przewodów), a nawet o takie dane, wydawałoby się odbiegające od zagadnienia awaryjności, jak produkcja, pobór czy straty wody.



Rys. 2. Sekwencja kosztów dla modeli drzew regresyjnych – przyłącza

Tabela 3. Dane węzłów dla modelu optymalnego

Przewody magistralne			
Nr węzła	Liczność	Średnia	Wariancja
1	8	0,125	0,001000
2	3	0,087	0,000089
3	5	0,148	0,000136
Przewody rozdzielcze			
1	8	0,346	0,001998
2	1	0,260	0,000000
3	7	0,359	0,001069
4	6	0,347	0,000256
6	3	0,333	0,000089
7	3	0,360	0,000067
5	1	0,430	0,000000
Przyłącza			
1	8	0,685	0,032275
2	1	0,260	0,000000
3	7	0,746	0,007396
4	6	0,715	0,002025
6	1	0,630	0,000000
7	5	0,732	0,000696
5	1	0,930	0,000000

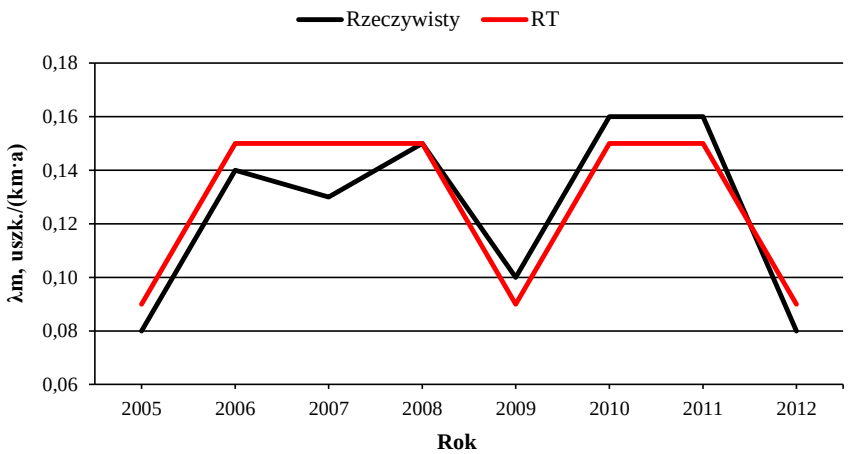
Wyniki przewidywania (z wykorzystaniem optymalnych modeli drzew regresyjnych) wskaźnika intensywności uszkodzeń przewodów magistralnych, rozdzielczych i przyłączy przedstawiono odpowiednio na rysunkach 3, 4 i 5 oraz w tabeli 4. Bardzo

ważnym elementem jest określenie tzw. ważności, czyli rankingu istotności predyktorów w skali 0–1. Takie podejście jest pomocne przy identyfikacji zmiennych posiadających istotną moc predykcyjną względem zmiennych zależnych [13]. W analizowanym zagadnieniu liczba uszkodzeń posiadała ważność równą 1 dla każdego rodzaju przewodu wodociągowego. Natomiast ważność zmiennej L wynosiła 0,33; 0,41 i 0,33 odpowiednio podczas przewidywania wskaźnika λ przewodów magistralnych, rozdzielczych i przyłączy.

Analiza tabeli 4 oraz rysunków 3–5 pokazuje, że rezultaty przewidywania wskaźnika awaryjności metodą drzew regresyjnych są zbieżne z wartościami rzeczywistymi. Struktury drzew regresyjnych (rys. 1) są relatywnie proste. Należy pamiętać, że do budowy tych modeli zastosowano jedynie dwie zmienne niezależne (L i N), co miało niewątpliwie wpływ na liczbę podziałów, gałęzi i węzłów. Im większy jest wektor predyktorów, tym bardziej rozbudowana architektura drzewa. Ponadto ważnym elementem (podobnie, jak w przypadku modelowania za pomocą innych metod regresyjnych) jest wiarygodna i pełna baza danych, bez zmiennych odstających. Konieczne wydaje się zatem, w kontekście ogólnie pojętej metodyki modelowania, dokonanie wstępnej analizy danych eksploatacyjnych i w razie potrzeby eliminacja zmiennych odstających. Jeśli odrzuconych danych jest zbyt wiele, stawia to pod znakiem zapytania zasadność modelowania. Niestety, rzeczywiste dane eksploatacyjne obarczone są niekiedy znacznymi błędami. W takich wypadkach należy rozsądnie podejść zarówno do odrzucania zmiennych odstających, jak i do włączania ich do analizy. Na tyle, na ile jest to możliwe należy dążyć (we współpracy z eksploatatorami) do wyjaśnienia nieścisłości i uzupełnienia brakujących parametrów.

Tabela 4. Wskaźnik intensywności uszkodzeń – modele optymalne drzew regresyjnych

Rok/typ przewodu	Przewody magistralne		Przewody rozdzielcze		Przyłącza	
	Wskaźnik intensywności uszkodzeń, uszk./(km·a)					
	Rzeczywisty	RT	Rzeczywisty	RT	Rzeczywisty	RT
2005	0,08	0,09	0,32	0,33	0,71	0,73
2006	0,14	0,15	0,43	0,43	0,93	0,93
2007	0,13	0,15	0,34	0,33	0,72	0,73
2008	0,15	0,15	0,34	0,33	0,74	0,73
2009	0,10	0,09	0,37	0,36	0,26	0,26
2010	0,16	0,15	0,35	0,36	0,71	0,73
2011	0,16	0,15	0,26	0,26	0,63	0,63
2012	0,08	0,09	0,36	0,36	0,78	0,73

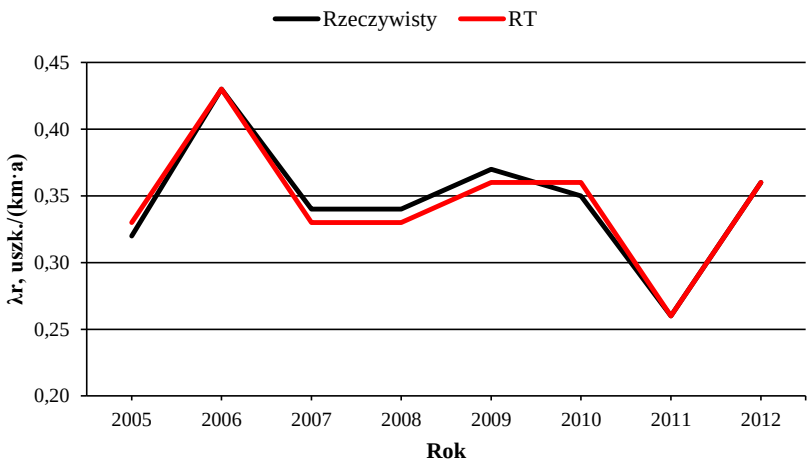


Rys. 3. Rzeczywiste i przewidywane wartości wskaźnika intensywności uszkodzeń przewodów magistralnych

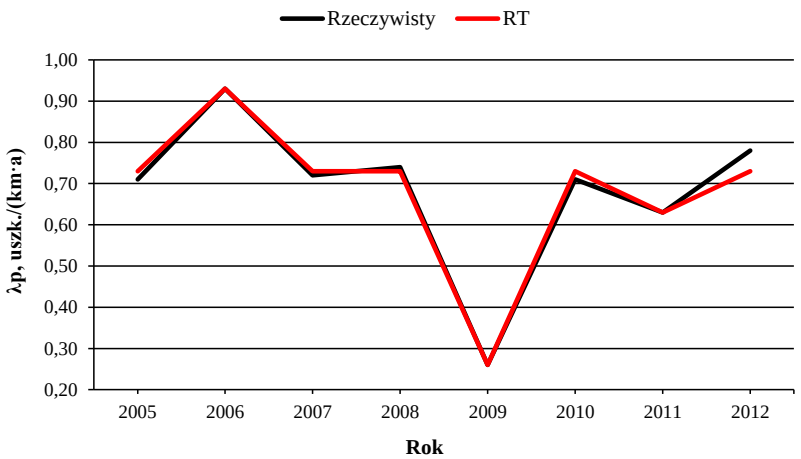
W przypadku przewodów magistralnych (rys. 3) wyniki uzyskane przy zastosowaniu optymalnego modelu RT są dobre. Rezultaty przewidywania są w latach 2005–2007 oraz 2012 nieco wyższe niż wartości rzeczywiste, natomiast w pozostałych latach nieco niższe. Jednak zauważalny jest właściwie identyczny trend zmian wskaźnika intensywności uszkodzeń, co symbolizują ułożone równolegle do siebie linie. Należy pamiętać, że na jakość przewidywania ma wpływ liczność próby. Zwłaszcza w odniesieniu do przewodów magistralnych, liczba lat (a co za tym idzie liczba przypadków wzorcowych) branych pod uwagę podczas tworzenia modelu RT, jest istotna z uwagi na relatywnie mniejszą liczbę występujących awarii w stosunku do uszkodzeń występujących na innych typach rurociągów. Sytuacja taka może mieć wpływ na uzyskanie nie do końca idealnej zbieżności, która charakteryzowana jest przez współczynniki korelacji (R) równy 0,939 oraz determinacji (R^2) równy 0,882.

Prognoza wskaźnika intensywności uszkodzeń przewodów rozdzielczych (rys. 4) jest bardzo dobra. Względny błąd przewidywania wyniósł maksymalnie trochę ponad 3%, co jest wynikiem satysfakcjonującym. W latach 2005, 2007–2010 oraz 2012 wskaźnik awaryjności wynosił ok. 0,3 uszk./(km·a). Przewidywanie tego wskaźnika metodą drzew regresyjnych dało dobre rezultaty. Należy zauważyć, że w latach 2006 i 2011 rzeczywiste wartości awaryjności znacznie odbiegały od danych z pozostałych lat. Również i w tym przypadku wyniki przewidywania pokryły się z wartościami rzeczywistymi. Jest to dowód na to, że algorytm RT jest w miarę uniwersalnym aproksymatorem. Inne metody regresyjne, np. sztuczne sieci neuronowe nie zawsze są w stanie dokonać właściwej prognozy danych nieco odstających od pozostałych wartości w zbio-

rze. Współczynnik korelacji Pearsona wyniósł 0,985, natomiast R^2=0,970, czyli zbieżność była lepsza niż w przypadku przewodów magistralnych, co właśnie może być związane z bardziej reprezentatywną próbką danych, o czym była już mowa.



Rys. 4. Rzeczywiste i przewidywane wartości wskaźnika intensywności uszkodzeń przewodów rozdzielczych



Rys. 5. Rzeczywiste i przewidywane wartości wskaźnika intensywności uszkodzeń przyłączy

Zaobserwowano praktycznie idealną zbieżność między rzeczywistą a przewidywaną awaryjnością przyłączy (rys. 5). Współczynniki R i R^2 były odpowiednio równe 0,993

i 0,986. Maksymalna wartość względnego błędu wyniosła 6,4% dla roku 2012. W pozostałych latach wartość ta wahała się w granicach 0–2,82%. Jednak analizując przydatność metod regresyjnych do przewidywania wskaźników awaryjności, należy brać pod uwagę nie tylko zgodność danych prognozowanych z rzeczywistymi, ale również typ rurociągu i jego wpływ na niezawodne działanie całej sieci dystrybucji wody. Awarie magistrali lub przewodu rozdzielczego mają donioślejsze skutki niż uszkodzenie nawet kilkunastu przyłączy domowych w jednym czasie. W związku z tym należy dążyć do zbudowania takiego modelu optymalnego, który będzie w sposób jeszcze dokładniejszy przewidywał wskaźnik intensywności uszkodzeń przewodów o średnicach większych, mających duży wpływ na działanie sieci wodociągowej w porównaniu np. do przyłączy domowych. Błędne oszacowanie wskaźnika λ w przypadku przewodów magistralnych i rozdzielczych (rys. 3 i 4, tab. 4) będzie miało większe konsekwencje z uwagi na wyższe koszty naprawy i tzw. koszty społeczne związane ze spadkiem ciśnienia w sieci lub chwilowymi przerwami w dostawie wody oraz innymi wydarzeniami mającymi związek z zaistniałą sytuacją awaryjną.

4. PODSUMOWANIE

W pracy zaprezentowano wyniki przewidywania, z wykorzystaniem metody drzew regresyjnych, wskaźnika intensywności uszkodzeń przewodów magistralnych, rozdzielczych i przyłączy w jednym z polskich miast. Tematyka pracy jawi się jako istotna z punktu widzenia prawidłowego i szybkiego szacowania poziomu niezawodności. Zbudowane modele RT mogą być przydatne w przypadku konieczności określenia wartości awaryjności na potrzeby np. podjęcia decyzji związanej z planowanymi remontami przewodów. Metodyka modelowania ukazana w niniejszej pracy jest pewnym novum w stosunku do dotychczasowego sposobu podejścia do przewidywania awaryjności przewodów wodociągowych. Przegląd literatury związanej z tą tematyką wykazał, że metoda drzew regresyjnych nie jest powszechnie stosowanym algorytmem podczas oceny poziomu niezawodności działania sieci wodociągowych. Fakt ten przyczynił się do podjęcia tego tematu. Zaprezentowane podejście i wyniki mają charakter wstępny, gdyż zmiennymi niezależnymi były podstawowe informacje: liczba uszkodzeń oraz długość danego typu rurociągu. Kolejnym etapem prac mogłaby być budowa bardziej skomplikowanych modeli drzew z wykorzystaniem większej liczby predyktorów, a w konsekwencji nawet stworzenie lasu losowego. Uzyskane w niniejszej pracy wyniki i zbieżność danych rzeczywistych z przewidywanymi (współczynniki determinacji na poziomie ok. 0,8 i 0,9) są, z inżynierskiego punktu widzenia, satysfakcjonujące. Względne błędy nie były wyższe niż 16%, 4% i 7% odpowiednio w przypadku przewodów magistralnych, rozdzielczych i przyłączy.

Należy pamiętać, że każde modelowanie obarczone jest błędem prognozy. Wybór modelu optymalnego powinien być związany nie tylko z uzyskaniem jak najlepszej zbieżności, ale także z oceną wpływu błędnego oszacowania. Skala i skutki awarii przewodów magistralnych lub rozdzielczych są nieporównanie większe niż uszkodzenia przyłączy domowych.

Praca została zrealizowana w ramach zlecenia B50519 z dotacji celowej przyznawanej dla Wydziału Inżynierii Środowiska Politechniki Wrocławskiej (W-7) przez Ministra Nauki i Szkolnictwa Wyższego na prowadzenie badań naukowych lub prac rozwojowych oraz zadań z nimi związanych służących rozwojowi młodych naukowców w latach 2015–2016.

LITERATURA

[1] BOGARDI I., FÜLÖP R., *A spatial probabilistic model of pipeline failures*, Periodica Polytechnica Civil Engineering, 2011, Vol. 55, No. 2, 161–168.
[2] CANDELIERI A., SOLDI D., CONTI D., ARCHETTI F., *Analytical leakages localization in water distribution networks through spectral clustering and support vector machines. The Icewater approach*, Procedia Engineering, 2014, Vol. 89, 1080–1088.
[3] HOTLOŚ H., *Ilościowa ocena wpływu wybranych czynników na parametry i koszty eksploatacyjne sieci wodociągowych*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2007.
[4] ILLA J.M.G., ALONSO J.B., MARRE M.S., *Nearest-Neighbours for time series*, Applied Intelligence, 2004, Vol. 20, No. 1, 21–35.
[5] IRIMIA-DIEGUEZ A.I., BLANCO-OLIVER A., VAZQUEZ-CUETO M.J., *A comparison of classification/regression trees and logistic regression in failure models*, Procedia Economics and Finance, 2015, Vol. 26, 23–28.
[6] KLEINER Y., RAJANI B., *Comprehensive review of structural deterioration of water mains: statistical models*, Urban Water, 2001, Vol. 3, No. 3, 131–150.
[7] KWIETNIEWSKI M., RAK J., *Niezawodność infrastruktury wodociągowej i kanalizacyjnej w Polsce*, Polska Akademia Nauk, Komitet Inżynierii Lądowej i Wodnej, Warszawa 2010.
[8] KUTYŁOWSKA M., *Modelling of failure rate of water-pipe networks*, Periodica Polytechnica Civil Engineering, 2015, Vol. 59, No. 1, 37–43.
[9] KUTYŁOWSKA M., ORŁOWSKA-SZOSTAK M., *Comparative analysis of water-pipe network deterioration – case study*, Water Practice and Technology, 2016, Vol. 11, No. 1, 148–156.
[10] Materiały udostępnione przez Przedsiębiorstwo Wodociągów i Kanalizacji.
[11] RAJANI B., KLEINER Y., *Comprehensive review of structural deterioration of water mains: physically based models*, Urban Water, 2001, Vol. 3, No. 3, 151–164.
[12] SCHEIDEGGER A., LEITAO J.P., SCHOLTEN L., *Statistical failure models for water distribution pipes – A review from unified perspective*, Water Research, 2015, Vol. 83, 237–247.
[13] Statistica 12.0, Electronic Manual.

REGRESSION TREES AS A TOOL FOR PREDICTION OF FAILURE FREQUENCY OF WATER PIPES

The paper shows the modelling, using regression trees method, of failure rate of water mains, distribution pipes and house connections in selected Polish city. Several models of regression trees were built. The choice of optimal models (separately for each type of conduit) was based on the cost analysis. The structure of regression tree contained independent variables (number of damages and the length of water pipes). Failure rate of three types of conduits was treated as dependent variable. Optimal models were characterized by the lowest costs and relatively simple architecture of tree. Operating data from years 2005–2012 were used for calculating the experimental values of failure rate and for regression trees models building. Optimal models for failure rate prediction of distribution pipes and house connections contained 3 divided nodes and 4 final nodes. On the other hand, regression tree for failure frequency modelling of water mains was less complicated in its structure and contained 1 divided node and 2 final nodes. The convergences between real and predicted values seem to be, from engineering point of view, satisfactory.