DOI: 10.2429/proc.2017.11(2)060 2017;11(2)

Małgorzata KUTYŁOWSKA[1]

DRZEWA REGRESYJNE JAKO NARZĘDZIE DO PROGNOZOWANIA WSKAŹNIKA INTENSYWNOŚCI USZKODZEŃ

REGRESSION TREES AS A TOOL FOR FORECASTING OF FAILURE RATE

Abstrakt: Przedstawiono możliwość zastosowania drzew regresyjnych (RT) do przewidywania wskaźnika intensywności uszkodzeń przewodów wodociągowych. Analiza wykorzystująca algorytm budowy drzew polega na znalezieniu zbioru logicznych warunków podziału oraz znalezieniu relacji pomiędzy predyktorami (zmiennymi niezależnymi) a zmienną zależną, co w konsekwencji prowadzi do uzyskania wyników prognozowania. Przewidywanie wskaźnika awaryjności przewodów rozdzielczych i przyłączy przeprowadzono na podstawie danych eksploatacyjnych z lat 2008-2014 dla jednej wybranej strefy zasilania w wodę średniej wielkości polskiego miasta. Zmiennymi niezależnymi były: długość danego typu przewodów oraz liczba uszkodzeń zaobserwowanych w danym roku na rurociągach rozdzielczych i przyłączach. Stworzono oddzielne modele drzew regresyjnych do modelowania awaryjności przewodów rozdzielczych i przyłączy. Obliczenia przeprowadzono w programie Statistica 13.1. Modele RT zarówno dla przyłączy, jak i przewodów rozdzielczych posiadały jeden węzeł dzielony i dwa końcowe. Wartość tzw. resubstytucji kosztów wynosiła 0,0056 i 0,00073 odpowiednio w modelu opisującym przyłącza i przewody rozdzielcze. Wyniki analiz i przewidywania pokazują, że drzewa regresyjne są dobrym narzędziem do przewidywania wskaźnika awaryjności przewodów wodociągowych, nawet stosując tak podstawowe predyktory, jak długość i liczba uszkodzeń.

Słowa kluczowe: metody regresyjne, przewody wodociągowe, prognozowanie, intensywność uszkodzeń

Wprowadzenie

Na obecnym etapie rozwoju, gdy znane są i wielokrotnie sprawdzone metody projektowania systemów wodociągowych oraz gdy hydraulika, a mówiąc ogólnie mechanika płynów, są dziedzinami poznanymi i stosowanymi, wydaje się, że nacisk w odniesieniu do systemów dystrybucji wody należy położyć na ich modernizację, właściwą eksploatację (uwzględniającą również stan jakości wody w odniesieniu do materiału, z jakiego wykonany jest przewód [1]) oraz badania niezawodności działania i prawidłowego zarządzania dostawą wody [2], które pozwolą na przedłużenie prawidłowego funkcjonowania infrastruktury podziemnej. Eksploatacja każdego z elementów systemu wodociągowego wymaga indywidualnego podejścia związanego z opisem i modelowaniem zjawisk zachodzących w danym obiekcie oraz uwzględniającego pełnioną funkcję. Dla przykładu, inaczej należy wykonywać badania awaryjności pompowni, które oparte mogą być np. o metodykę propagacji fali z zastosowaniem modelu Lagrange'a [3], a zupełnie w inny sposób podchodzi się do detekcji online nieszczelności na sieci wodociągowej lub sposobu typowania przewodów wodociągowych do remontu (np. z wykorzystaniem sztucznej inteligencji) [4, 5]. Problematyka niezawodności działania i awaryjności sieci wodociągowych, a także badania związane ze stratami wody były i są przedmiotem analiz wielu zespołów badawczych w Polsce [6-9], co przyczyniło się do

[1] Wydział Inżynierii Środowiska, Politechnika Wrocławska, ul. Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław, tel. 71 320 40 84 , email: malgorzata.kutylowska@pwr.edu.pl
Praca była prezentowana podczas konferencji ECOpole'17, Polanica Zdrój, 4-7.10.2017

rozwinięcia tej dziedziny nauki również w kontekście modelowania nie tylko w Polsce, ale także za granicą [10, 11]. Niniejsza praca jest zatem próbą uzupełnienia dotychczasowych badań o aspekt modelowania wskaźników niezawodnościowych (na przykładzie wskaźnika intensywności uszkodzeń przewodów wodociągowych) z wykorzystaniem metod regresyjnych (tzw. metod uczenia maszyn). Do takich metod można zaliczyć drzewa regresyjne i klasyfikacyjne, które znajdują szerokie zastosowanie jako narzędzie do modelowania w wielu dziedzinach, jednak właściwie do tej pory brak jest w światowej i krajowej literaturze informacji o wykorzystaniu tej metodyki do analizy poziomu awaryjności i prognozowania wskaźnika intensywności uszkodzeń przewodów wodociągowych, co skłoniło do podjęcia właśnie tego tematu.

Metodyka badań

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 [12]. Ogólnie rzecz ujmując, drzewo regresyjne (RT) lub klasyfikacyjne (CT) jest grafem skierowanym, zawierającym korzeń i węzły (liście), w których sprawdzane są warunki dotyczące zmiennych, a także gałęzie zawierające reguły decyzyjne. Metoda drzew regresyjnych jest z reguły łatwiejsza w implementacji i analizie wyników niż metoda drzew klasyfikacyjnych [12]. Analiza wykorzystująca algorytm budowy drzew polega na znalezieniu zbioru logicznych warunków podziału oraz znalezieniu relacji pomiędzy predyktorami a zmienną zależną, co w konsekwencji prowadzi do uzyskania wyników przewidywania [12]. Zaletą stosowania drzew jest relatywnie prosta interpretacja wyników oraz dobre rezultaty predykcji [13]. Ponadto cechą modeli drzew regresyjnych jest ich odporność na dane odstające, które niestety mogą się często i z różnych przyczyn pojawić w danych eksploatacyjnych uzyskiwanych z przedsiębiorstw wodociągowych. W przypadku pojawienia się danych odstających są one izolowane w małych węzłach. Jeśli wartości tych jest niewiele, to mogą być one pominięte [12]. 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.

Przewidywanie wskaźnika intensywności uszkodzeń (λ [uszk./(km·rok)]) przewodów wodociągowych przeprowadzono z wykorzystaniem metody drzew regresyjnych. Dokonano oddzielnie predykcji wskaźnika awaryjności przewodów rozdzielczych (λ_r) i przyłączy domowych (λ_p), co oznaczało konieczność budowy dwóch różnych modeli drzew. Wskaźniki λ były zmiennymi zależnymi, natomiast predyktorami (zmiennymi niezależnymi) były: długość danego typu przewodu (L_p i L_r) oraz liczba uszkodzeń (N_p i N_r) zarejestrowana w danym roku odpowiednio dla przyłączy i rurociągów rozdzielczych. Do analizy specjalnie zostały wybrane tak podstawowe zmienne, jak długość i liczba uszkodzeń. Wynika to z faktu, iż tego typu dane są na pewno notowane w przedsiębiorstwach wodociągowych, a zatem są łatwo dostępne. Ponadto w niniejszej pracy sprawdzono, czy właśnie tak podstawowe informacje o rurociągach (bez wnikania w szczegóły dotyczące materiału i średnicy rury) są przydatne do przewidywania wskaźnika intensywności uszkodzeń za pomocą drzew regresyjnych.

Dane eksploatacyjne, uzyskane z przedsiębiorstwa wodociągowego w jednym z polskich średniej wielkości miast z lat 2008-2014, posłużyły do wyznaczenia rzeczywistego wskaźnika λ oraz do przewidywania wskaźnika awaryjności za pomocą metody drzew regresyjnych. Cały system dystrybucji wody podzielony jest na 55 stref zasilania. W niniejszej analizie skupiono się na jednej wybranej strefie zasilania, w której ciśnienie wewnątrz rurociągów wynosiło 0,4 MPa. Wartości eksperymentalnych zmiennych zależnych i predyktorów w latach 2008-2014 zestawiono w tabeli 1. W ciągu 7 lat eksploatacji całkowita długość przewodów rozdzielczych i przyłączy w analizowanej strefie nie ulegała zmianie.

Tabela 1

Zmienne zależne i predyktory

Table 1

Dependent variables and predictors

L_r [km]	L_p [km]	N_r [uszk.]	N_p [uszk.]	λ_r [uszk./(km·rok)]	λ_p [uszk./(km·rok)]
17,5	14,2	2-5	3-10	0,11-0,29	0,21-0,70

Obliczenia zaprezentowane w niniejszej pracy przeprowadzono w programie Statistica 13.1.

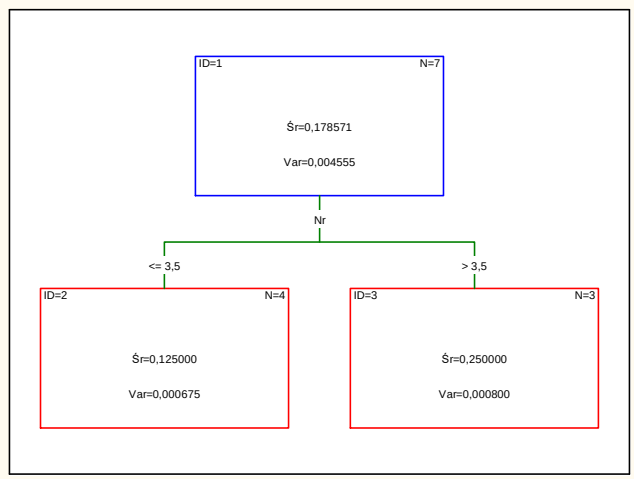
Wyniki badań i ich dyskusja

Zbudowano kilka modeli drzew regresyjnych do przewidywania wskaźników λ_r i λ_p. Wybrano optymalne modele RT, których struktury przedstawiono na rysunku 1. Wybór modelu najbardziej odpowiedniego do przewidywania poziomu awaryjności rurociągów rozdzielczych i przyłączy dokonany został z uwzględnieniem takich parametrów, jak: najmniejsze tzw. koszty resubstytucji, niewielka złożoność modelu, a także jakość przewidywania, czyli zbieżność rzeczywistej (eksperymentalnej) zmiennej zależnej z wartościami uzyskanymi podczas modelowania.

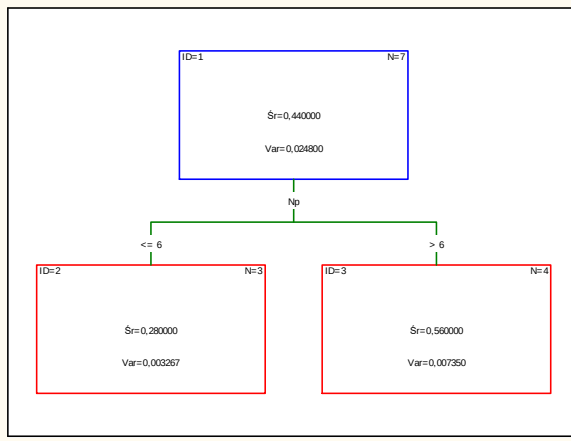
Pozornie mogłoby się wydawać, że architektury obu modeli (rys. 1) są takie same. Rzeczywiście liczba węzłów dzielonych (1 węzeł) i węzłów końcowych (2 węzły) jest taka sama, ale z uwagi na inne wartości zmiennych niezależnych wartości średniej i wariancji w danym węźle różnią się znacznie (rys. 1a i 1b). Ponadto inna wartość odpowiedzialna jest za podział węzła dzielonego na dwa węzły końcowe. Do oceny jakości modelu posłużono się pojęciem tzw. resubstytucji kosztu. W metodzie RT [13] jest to uogólnienie 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. W przypadku modelu RT, opisującego wskaźnik intensywności uszkodzeń przewodów rozdzielczych, koszt wyniósł 0,00073, natomiast dla modelu charakteryzującego przyłącza wodociągowe wartość ta była o rząd większa i równała się 0,0056. Dla pozostałych modeli (innych niż model optymalny) koszty rosły liniowo wraz ze zwiększaniem się liczby węzłów dzielonych i końcowych (czyli wraz ze zwiększaniem się złożoności modelu), zarówno dla rurociągów rozdzielczych, jak i przyłączy. Porównanie dwóch wartości kosztów w wybranych modelach RT do prognozowania poziomu awaryjności wskazuje, że pomimo

takiej samej architektury drzewa regresyjnego jakość modelowania, definiowana jako m.in. wartość resubstytucji kosztu, jest znacząco inna.

a)



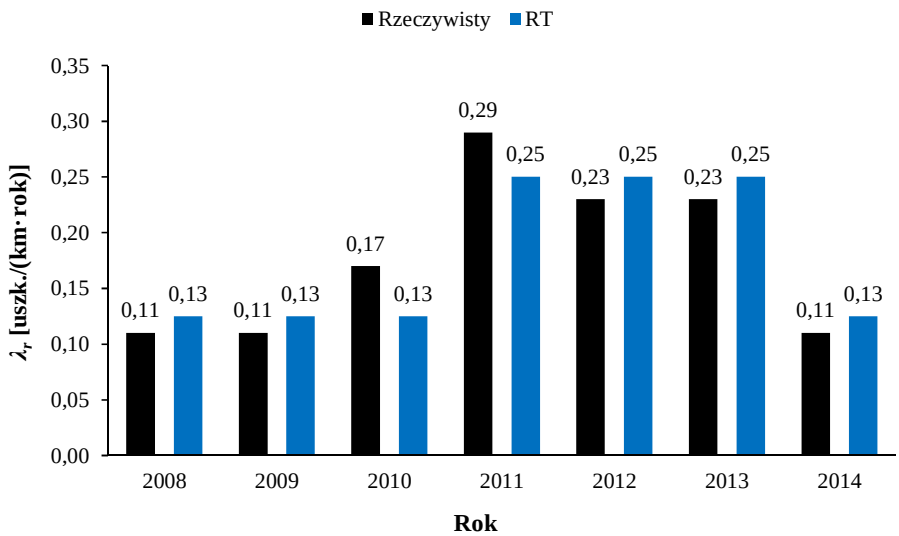
b)



Rys. 1. Optymalne struktury drzewa regresyjnego: a) przewody rozdzielcze, b) przyłącza domowe
Fig. 1. Optimal regression tree structures: a) distribution pipes, b) house connections

Istotnym elementem, charakteryzującym model i opisującym wpływ zmiennych niezależnych na wartość zmiennej zależnej, 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 [12, 13]. Dla dwóch modeli optymalnych liczba uszkodzeń w danym roku (N_p i N_r) była zmienną niezależną, której tzw. ważność była największa i wynosiła 1. Jest to dodatkowo uwidocznione na rysunku 1, gdyż zmienną odpowiedzialną za podział na kolejne poziomy drzewa regresyjnego jest właśnie liczba uszkodzeń. Z przeprowadzonej analizy wynika, że

długość przewodów nie ma właściwie wpływu na budowę, a następnie na jakość modeli RT. Być może wynika to z faktu, że wektor predyktorów nie był liczny, tzn. składał się tylko z dwóch zmiennych. Zupełnie inna sytuacja obserwowana była podczas analizy wyników modelowania za pomocą RT wskaźnika awaryjności w innym polskim mieście [14]. W tym przypadku [14] długość rurociągów była zmienną dominującą. Jednakże może wiązać się to z zupełnie innym wektorem zmiennych niezależnych wykorzystanych do budowy modelu. Poza długością predyktorami we wspomnianej pracy były takie zmienne, jak: średnica, rok budowy i materiał. Niemniej jednak w niniejszej pracy celowo, należy to podkreślić, wykorzystano tak podstawowe zmienne niezależne, jak *N* i *L*, aby sprawdzić możliwość stosowania drzew regresyjnych do modelowania poziomu awaryjności przewodów wodociągowych nawet w przypadku posiadania niewielu informacji na temat rozpatrywanego systemu dystrybucji wody.

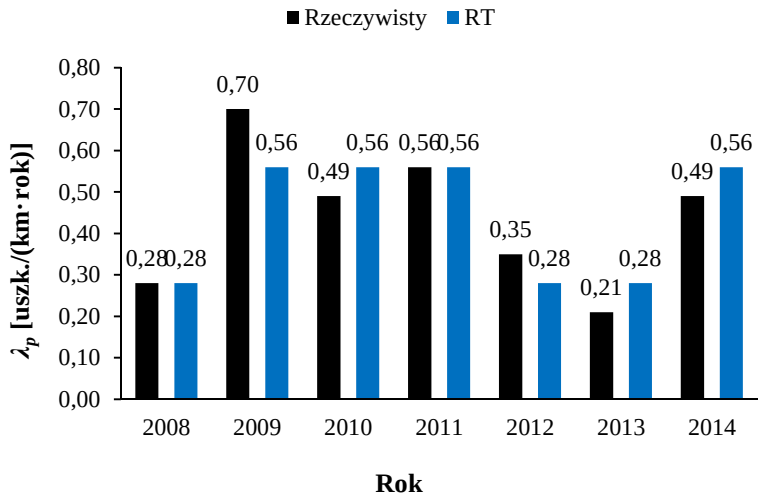


Rys. 2. Rzeczywiste i przewidywane wartości wskaźnika intensywności uszkodzeń (λ_r) przewodów rozdzielczych
Fig. 2. Real and predicted values of failure rate (λ_r) of distribution pipes

Analiza wyników przewidywania (rys. 2 i 3) wskaźnika intensywności uszkodzeń przewodów rozdzielczych i przyłączy domowych wskazuje, że zastosowanie tylko dwóch zmiennych w wektorze predyktorów nie wpłynęło na jakość modelowania. Wyniki nie są może idealnie zbieżne, jak to uzyskano w pracy [14], ale z inżynierskiego punktu widzenia są one satysfakcjonujące. W przypadku rurociągów rozdzielczych maksymalny bezwzględny błąd wyniósł 0,04 uszk./(km·rok), a podczas modelowania awaryjności przyłączy domowych 0,14 uszk./(km·rok). Wyniki przedstawione na rysunku 2 wskazują, że dla pięciu z siedmiu analizowanych lat wartość wskaźnika intensywności uszkodzeń jest nieco większa w przypadku modelowania w stosunku do wartości rzeczywistych. W odniesieniu do przewodów rozdzielczych tego typu niewielkie przeszacowanie nie budzi

obaw o jakość modelowania, gdyż wyższy przewidywany poziom awaryjności w stosunku do rzeczywistego może jedynie podziałać mobilizująco na eksploatatora sieci i skłonić do podjęcia wcześniejszej decyzji o renowacji lub wymianie wybranych odcinków sieci.

Należy pamiętać, że ranga rurociągów rozdzielczych jest większa niż przyłączy domowych, co sprawia, że niedoszacowanie wskaźnika intensywności uszkodzeń w latach 2009 i 2012 (rys. 3) nie może być raczej powodem do stwierdzenia, że modele RT nie są dobrym narzędziem do przewidywania poziomu awaryjności przyłączy. Jednakże warto wspomnieć, że wyniki przedstawione w niniejszej pracy dotyczą przewidywania wskaźników λ_r i λ_p tylko na próbie uczącej, czyli, innymi słowy, na próbie danych zastosowanych do budowy modelu. W związku z tym zasadne wydaje się dalsze pogłębianie wiedzy i metodyki badań w zakresie poznania sposobów budowy, a następnie zastosowania stworzonych już modeli drzew regresyjnych tak, aby można w dalszym etapie sprawdzić ich jakość na zmiennych niezależnych niewłączonych wcześniej do analizy.



Rys. 3. Rzeczywiste i przewidywane wartości wskaźnika intensywności uszkodzeń (λ_p) przyłączy

Fig. 3. Real and predicted values of failure rate (λ_r) of house connections

Wnioski

Modelowanie i przewidywanie wskaźnika intensywności uszkodzeń przewodów wodociągowych wydaje się, na obecnym etapie badań dotyczących niezawodności działania systemów komunalnych, zagadnieniem niezwykle istotnym z uwagi na konieczność podejmowania szybkich decyzji w sytuacjach wystąpienia poważniejszych awarii. Istnieje wiele metod modelowania, jednak w ostatnim czasie coraz częściej stosowane są tzw. metody uczenia maszyn, do których zaliczyć można również metodę drzew regresyjnych. W niniejszej pracy zastosowano modele RT do przewidywania wskaźnika awaryjności rurociągów rozdzielczych i przyłączy. Modele RT zarówno dla przyłączy, jak i przewodów rozdzielczych posiadały jeden węzeł dzielony i dwa końcowe. Wartość tzw. resubstytucji kosztów wynosiła 0,0056 i 0,00073 odpowiednio w modelu

opisującym przyłącza i przewody rozdzielcze. Wyniki, pomimo zastosowania podstawowych zmiennych niezależnych i bardzo prostej architektury drzewa, są satysfakcjonujące i świadczą o możliwości stosowania metody RT jako podejścia alternatywnego do innych sposobów modelowania. Niemniej jednak konieczne wydają się dalsze badania w zakresie możliwości aplikacji modeli RT z uwzględnieniem wektora zmiennych niezależnych niewłączonych wcześniej do budowy modelu. Badania takie będą w najbliższym czasie prowadzone w odniesieniu do innych systemów dystrybucji wody w celu uzyskania wyników pozwalających na dokonanie pewnych uogólnień i umożliwiających postawienie dalej idących tez dotyczących przewidywania poziomu awaryjności i niezawodności działania infrastruktury podziemnej.

Podziękowania

Pracę zrealizowano w ramach działalności statutowej Wydziału Inżynierii Środowiska Politechniki Wrocławskiej, finansowanej ze środków Ministerstwa Nauki i Szkolnictwa Wyższego w latach 2017-2018 roku, nr projektu 0401/0006/17.

Literatura

[1] Musz A, Kowalska B. Ecol. Chem. Eng S. 2015;22(2):219-229. DOI: 10.1515/eces-2015-0012.
[2] Pietrucha-Urbanik K, Studziński A. Ecol Chem Eng A. 2016;23(3):299-311. DOI: 10.2428/ecea.2016.23(3)25.
[3] Meniconi S, Brunone B, Ferrante M, Capponi C, Carrettini CA, Chiesa C, et al. J. Hydroinform. 2015;17(3):377-389. DOI: 10.2166/hydro.2014.038.
[4] Mounce SR, Boxall JB, Machell J. J Water Res. Pl-ASCE 2010;136(3):309-318. DOI: 10.1061/(ASCE)WR.1943-5452.0000030#sthash.Mt3pYumS.dpuf.
[5] Kamiński K, Kamiński W, Mizerski T. Proc ECOpole. 2016;10(2):661-666. DOI: 0.2429/proc.2016.10(1)072.
[6] Kowalski D, Miszta-Kruk K. Eng Failure Analysis. 2013;35:736-742. DOI: 10.1016/j.engfailanal.2013.07.017.
[7] Kwietniewski M, Rak J. Niezawodność infrastruktury wodociągowej i kanalizacyjnej w Polsce [Reliability of water supply and wastewater disposal infrastructure in Poland]. Warszawa: Komitet Inżynierii Lądowej i Wodnej PAN; 2010.
[8] Tchórzewska-Cieślak B. Environ Prot Eng. 2009;35(2):29-35. http://epe.pwr.wroc.pl/2009/Tchorzewska_2-2009.pdf.
[9] Musz-Pomorska A, Iwanek M, Parafian K, Wójcik K. E3S Web of Conferences 17, 00062 (2017). DOI: 10.1051/e3sconf/20171700062.
[10] Iwanek M, Suchorab P, Karpińska-Kiełbasa M. Periodica Polytechnica Civ Eng. Online first. Volume (2017), paper 9728. DOI: 10.3311/PPci.9728.
[11] Francis RA, Guikema SD, Henneman L. Reliab Eng Syst Safe. 2014;130:1-11. DOI: 10.1016/j.ress.2014.04.024.
[12] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Boca Raton, USA: Chapman Hall/CRC; 1984. ISBN: 9780412048418.
[13] Statistica 13.1. Electronic Manual. https://www.statsoft.pl/textbook/stathome.html.
[14] Kutyłowska M. E3S Web of Conferences. 2017;22;00097. DOI: 10.1051/e3sconf/20172200097.

REGRESSION TREES AS A TOOL FOR FORECASTING OF FAILURE RATE

Faculty of Environmental Engineering, Wrocław University of Science and Technology, Wrocław

Abstract: This paper presents the possibility of applying regression trees (RT) to predict the failure rate of water pipes. An analysis using a tree building algorithm consists in finding a set of logical division conditions, and relations between the predictors (independent variables) and the dependent variable, which leads to prediction results. The failure rate of distribution pipes and house connections was predicted on the basis of operational data for the years 2008-2014 in one selected zone in medium sized Polish city. Independent variables were such parameters as length of conduits and number of damages registered in each year at distribution pipes and house connections. Models for failure rate forecasting of distribution pipes and house connections were created separately. The calculations were carried out using the Statistica 13.1 software. RT models for house connections and distribution pipes have one divided node and two end nodes. The value of resubstitution cost amounted to 0.0056 and 0.00073 for models describing house connections and distribution pipes, respectively. The results of analysis and forecasting investigations show that regression trees are relatively good tool for failure rate prediction, using even such basic predictors as length and number of damages.

Keywords: regression methods, water pipes, prediction, failure intensity