Imputacja brakujących danych na potrzeby modelowania sieci ciepłowniczych

Imputation of missing data for use in district heating networks modeling

JAKUB KUŚ, MICHAŁ ŻURAWSKI, ŁUKASZ MIKA

DOI: 10.17512/INSTAL.2026.08.02

W ramach pracy podjęto problematykę uzupełniania brakujących danych charakteryzujących działanie węzłów cieplnych, w sytuacji niedostępności dokładnych danych telemetrycznych. Kompletne zbiory danych wejściowych są niezbędne do zasilenia modelu obliczeniowego sieci ciepłowniczej. Scharakteryzowano i porównano trzy metody imputacji: imputację średnią, metodę typu hot-deck oraz imputację za pomocą regresji liniowej. Do oceny metod wykorzystano rzeczywiste dane telemetryczne, w których w sposób losowy wygenerowano luki. Najskuteczniejsza w uzupełnianiu braków okazała się metoda regresji liniowej, osiągając wskaźnik dopasowania R^2 mocy węzłów cieplnych na poziomie do 0,95 i natężenia przepływu na poziomie do 0,86. W artykule wskazano źródła potencjalnie wartościowej wiedzy pomocniczej do wykorzystania w celu uzupełniania brakujących danych.
Słowa kluczowe: sieci ciepłownicze, modelowanie hydrauliczne, imputacja, brakujące dane

This study addresses the issue of imputing missing data relating to the operation of district heating substations when accurate telemetry data is unavailable. A complete set of input data is required to inform a mathematical model of a district heating network. Three imputation methods were characterized and compared: mean imputation; the hot-deck method; and imputation using linear regression. The methods were evaluated using actual telemetry data in which missing values were randomly induced. Linear regression proved to be the most effective method for filling the gaps, achieving R^2 scores of up to 0.95 for substation heat power and up to 0.86 for volumetric flow rate. The article identifies sources of potentially valuable auxiliary knowledge that could be used to fill in missing data.
Keywords: district heating networks, hydraulic modelling, imputation, missing data

Wprowadzenie

Do wykonania obliczeń rozpływów wody i strat ciśnienia w sieci ciepłowniczej konieczne jest dysponowanie pełnym zbiorem niezbędnych danych wejściowych [1]. Jedną z części wymaganego zbioru danych są informacje na temat przepływów i spadków temperatury w węzłach cieplnych, które mogą zostać pozyskane na podstawie archiwalnych danych z systemów telemetrycznych. Przeszkodą w określeniu tych parametrów może być niepełne pokrycie odbiorców ciepła systemami telemetrycznymi albo awarie lub błędy związane z przetwarzaniem, transmisją i archiwizacją danych pomiarowych [2]. Tego rodzaju problemy mogą prowadzić do sytuacji, że dla pewnej podgrupy węzłów cieplnych nie istnieje żaden, nawet mały, zbiór danych pozwalających bezpośrednio wyznaczyć parametry niezbędne do zasilenia modelu cieplno-hydraulicznego sieci ciepłowniczej.

Wówczas pojawia się kwestia, w jaki sposób poradzić sobie z niekompletnością danych. Podstawy teorii danych brakujących i zasadnicze metody radzenia sobie z tym problemem wywodzą się m.in. z dziedziny nauk społecznych i pokrewnych, które często operują na danych niepełnych, np. ze względu na unikanie z różnych przyczyn przez respondentów odpowiedzi na niektóre pytania z kwestionariuszy [3] lub pracę na zasobach danych z założenia niekompletnych, np. dlatego, że trudno byłoby je pozyskać w całości [4]. Braki w danych są jednak istotnym problemem interdyscyplinarnym [2,5]. W zbiorach danych pozyskiwanych z systemów energetycznych, w tym ciepłowniczych, często występującym przykładem problemu związanego z niekompletnością danych są brakujące wartości w szeregach czasowych danych pomiarowych. Współcześnie niekompletność danych w ujęciu teoretycznym opisuje się jako skutek działania fenomenu probabilistycznego, prowadzącego do wyodrębnienia się w zbiorze danych dwóch części: obserwowanej, na którą składają się znane wartości zmiennych i części nieobserwowanej, złożonej ze zmiennych o nieznanych wartościach.

Systematyczne uzupełnienie danych brakujących konkretnymi wartościami, w taki sposób, aby otrzymać kompletny zbiór danych, nazywa się imputacją [6, 7]. Według [4] informację wprowadzoną w ten sposób, czyli imputowaną na podstawie innej pomocniczej wiedzy, nazywa się implantem statystycznym.

Imputacja może zostać przeprowadzona różnymi metodami. Prawdopodobnie najprostszą z nich jest imputacja średnią. Polega ona na zastąpieniu brakujących danych wartością średnią rozważanego parametru, wyznaczoną z zasobu danych obserwowanych. Inną prostą metodą jest imputacja hot-deck, polegająca na zastępowaniu brakujących danych wartością zaczerpniętą z zasobu danych obserwowanych od obiektu „podobnego" –

mgr inż. Jakub Kuś ORCID: 0009-0006-9309-4844, Autor korespondencyjny: e-mail: jkus@agh.edu.pl, mgr inż. Michał Żurawski ORCID: 0009-0004-4052-7322, Akademia Górniczo-Hutnicza, Wydział Energetyki i Paliw, Al. Mickiewicza 30, 30-059 Kraków;
dr hab. inż., Łukasz Mika, prof. AGH ORCID: 0000-0002-0750-7694, Akademia Górniczo-Hutnicza, Wydział Energetyki i Paliw, Katedra Maszyn Cieplnych i Przepływowych, Al. Mickiewicza 30, 30-059 Kraków

Data zgłoszenia: 14.06.2026. Data przyjęcia: 16.07.2026. Data publikacji: 26.08.2026.

„dawcy" [3]. W zależności od potrzeb to „podobieństwo" może zostać zdefiniowane na różne sposoby [4], np. za obiekt podobny może zostać uznany ten, który charakteryzuje się najbliższą wartością jednej z obserwowanych zmiennych. Kolejną metodą imputacji jest regresja. W ramach tej metody buduje się model regresji, np. liniowej, na podstawie obserwowanej części zasobu danych, a następnie stosuje się ten model do części nieobserwowanej. Wprowadzając do tej deterministycznej metody element losowości, uzyskuje się stochastyczną imputację regresyjną [3].

Powyższe metody opierają się na założeniu, że część nieobserwowana, która zostanie imputowana, zależy od obserwowanej części zasobu danych. W przypadku danych pomiarowych takie założenie często jest uzasadnione, natomiast nie zawsze zmienne imputowane da się opisać bezpośrednią zależnością funkcyjną względem zmiennych obserwowanych [6]. W przypadku niemożności określenia takiej bezpośredniej zależności funkcyjnej, możliwą do zastosowania klasą metod są metody opierające się na modelowaniu hipotetycznego rozkładu prawdopodobieństwa rozważanych zmiennych nieobserwowanych na podstawie parametrów zbioru zmiennych obserwowanych [2,6].

Aplikacyjne zastosowanie metod imputacji można ogólnie podzielić na imputację jednostkową (pojedynczą), w której każda brakująca wartość jest uzupełniana jednokrotnie oraz imputację wielokrotną, w której każda brakująca wartość uzupełniana jest kilkukrotnie, a dalsza analiza statystyczna kolejnych wariantów uzupełnionego zbioru danych jest ostatecznie łączona w jeden końcowy wynik [7,8]. Dobre rezultaty można uzyskać, stosując metody oparte na uczeniu maszynowym [9], przy czym tego typu techniki wymagają dysponowania dużym zasobem danych obserwowanych, aby odpowiednio wytrenować model. Szerszy opis wyżej wymienionych oraz innych metod imputacji można znaleźć w przeglądach tematycznych, m.in. [10, 11].

Na gruncie zagadnień związanych z dystrybucją ciepła i pokrewnych, w literaturze naukowej obecne są przykłady wykorzystania metod imputacyjnych w celu dopełnienia danych pomiarowych (w postaci szeregów czasowych) pochodzących z ciepłomierzy, (uzupełnienie w ten sposób interwałów czasowych, w których wystąpił brak danych) [12-14]. W pracy [12] porównano wyniki imputacji różnymi metodami brakujących danych pomiarowych ze zbioru około 3000 ciepłomierzy. Udział brakujących obserwacji w tym zbiorze wynosił około 0,3% wszystkich rekordów. W artykule [13] udział braków w źródłowym zasobie danych był większy i wynosił około 3,7%. W pracy [14] w przypadku wartości chwilowych najlepiej sprawdziła się metoda imputacji oparta na średniej ruchomej, a w przypadku imputacji wartości skumulowanych najbardziej skuteczna okazała się interpolacja liniowa. W publikacji [2] zastosowano różne metody imputacji brakujących danych z liczników energii elektrycznej i pokazano, że skuteczność imputacji malała wraz z rosnącym udziałem braków. W pracy [15] zademonstrowano wykorzystanie imputacji wielokrotnej do uzupełnienia braków w danych pomiarowych z liczników energii elektrycznej. W artykule [16] przedstawiono przykład podejścia mieszanego, integrującego różne metody imputacji.

Definicja problemu

W powyżej wymienionych przykładach z literatury naukowej imputacja wykorzystywana była głównie do uzupełniania luk w profilach czasowych danych z liczników. Przedmiotem niniejszego artykułu będzie sytuacja, w której dla pewnej grupy węzłów cieplnych w ogóle nie są dostępne godzinowe dane telemetryczne pozwalające na bezpośrednie wyznaczenie przepływów i spadków temperatury, stanowiących niezbędne dane wejściowe dla modelu cieplno-hydraulicznego sieci ciepłowniczej. W danym punkcie pracy sieci powyższe parametry w każdym z węzłów cieplnych związane są zależnościami:

$$\dot{Q} = \rho \dot{V} c_p \Delta T \tag{1}$$

$$\dot{Q} = \varphi \dot{Q}_{nom} \tag{2}$$

gdzie:
$\dot{Q}$ – moc cieplna węzła [kW], ρ – gęstość czynnika [kg/m^3], $\dot{V}$ – przepływ objętościowy czynnika przez węzeł [m^3/s], c_p – ciepło właściwe czynnika [kJ/(kg · °C)], ΔT – spadek temperatury na węźle [°C], φ – współczynnik wykorzystania mocy zamówionej dla danego węzła [-], $\dot{Q}_{nom}$ – moc zamówiona węzła cieplnego [kW].

Znając dwie wartości spośród: mocy, strumienia przepływu i spadku temperatury, za pomocą wzoru (1) można obliczyć trzeci z parametrów.

Zadanie rozważone w niniejszej pracy polega na imputowaniu na podstawie innej wiedzy pomocniczej (np. wiadomej części danych pomiarowych, danych GIS lub technicznych), wartości dwóch z trzech powyższych parametrów dla podzbioru węzłów cieplnych, w których nie jest znany żaden z tych parametrów, w celu uzyskania kompletnego zbioru danych wejściowych do zasilenia modelu sieci ciepłowniczej. Odzwierciedla to przykładową sytuację, w której w danym systemie ciepłowniczym pewna liczba węzłów cieplnych jest objęta systemem telemetrii i parametry tych węzłów potrzebne do modelowania sieci ciepłowniczej można wyznaczyć bezpośrednio z danych telemetrycznych, a pewna grupa węzłów nie jest objęta systemem telemetrii i parametry tych węzłów potrzebne do modelowania sieci ciepłowniczej, które nie mogą zostać określone bezpośrednio z danych telemetrycznych, muszą zostać wyznaczone w inny sposób.

Głównym celem pracy jest porównanie, jak sprawdzają się w takim zastosowaniu trzy metody oparte na imputacji: średnią, hot-deck i regresją. W obliczeniach posłużono się zasobem danych pochodzącym z fragmentu rzeczywistego systemu ciepłowniczego, obejmującego 42 węzły cieplne. Wśród odbiorców ciepła znajdują się budynki mieszkalne w zabudowie wielo- i jednorodzinnej, obiekty usługowe, biurowe i przemysłowe. Zasób danych obejmuje kompletne dane z telemetrii wszystkich węzłów cieplnych w rozdzielczości godzinowej (moce, przepływy i spadki temperatur) z okresu od 1 stycznia do 14 lutego 2024 r. W obliczeniach wykorzystano również pomocnicze zasoby danych innego typu, wymaganych stosownie do metody (np. dane GIS).

Powyższe dane telemetryczne przekształcono, pozyskując cztery zbiory danych bazowych:
- trzy zbiory uśrednione A_n10, A_0 i A_p10, odpowiadające uśrednionym parametrom węzłów cieplnych przy temperaturach zewnętrznych bliskich odpowiednio –10°C, 0°C i +10°C;
- jeden zbiór B_0, odpowiadający chwilowym parametrom węzłów cieplnych w jednym z kilkugodzinnych interwałów czasowych, gdy działanie sieci ciepłowniczej było bardzo stabilne i zbliżone do stanu stacjonarnego.

Wykorzystanie powyższych zbiorów danych bazowych („A" – stany uśrednione, „B" – stan chwilowy, „quasi stacjonarny") ma na celu umożliwić sprawdzenie skuteczności imputacji dla danych reprezentujących różne stany sieci.

W powyższych zbiorach danych bazowych sztucznie wygenerowano braki na poziomie 10, 20 lub 30 losowo wybranych węzłów. Brakujące dane imputowano każdą z rozważanych metod: średnią, hot-deck i regresją, w oparciu o obserwowaną część zbioru danych i, stosownie do analizowanej metody, dane pomocnicze. Takie podejście umożliwia sprawdzenie zarówno skuteczności poszczególnych metod, jak i wpływu poziomu brakujących danych.

Powyższy ciąg operacji (indukcję braków i następcze ich uzupełnienie implantami statystycznymi za pomocą trzech metod) powtórzono dla każdej kombinacji wariantów (dane bazowe × poziom wyindukowanych braków) po 1000 razy, aby umożliwić ocenę rozproszenia uzyskiwanych wyników, powodowanego przez zróżnicowanie obserwowanej części zbiorów danych na skutek losowej indukcji braków. Takie podejście umożliwia ogólne wnioskowanie na temat odporności metody na niereprezentatywność próbki – obserwowanej części zbioru danych.

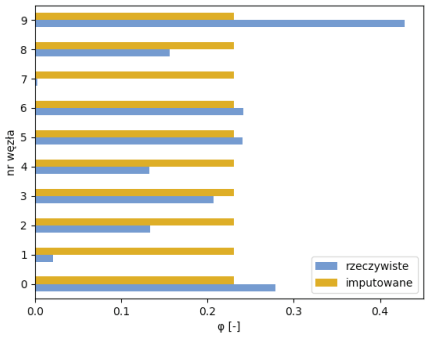
W toku dalszych rozważań sekwencja losowej indukcji braków i następczego uzupełnienia implantami statystycznymi będzie nazywana iteracją, a część obserwowana zasobu danych w ramach pojedynczej iteracji, stanowiąca dopełnienie części nieobserwowanej, będzie nazywana próbką. Losowo wyindukowana w danej iteracji część nieobserwowana zasobu danych będzie nazywana przeciwpróbką.

Metody imputacji

Imputacja średnią

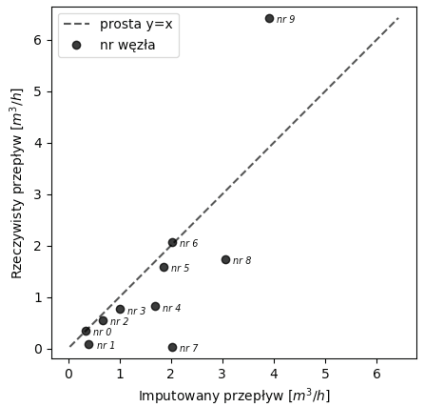
W ramach tej metody brakujące dane w części nieobserwowanej zasobu danych (przeciwpróbce) uzupełniano wartościami średnimi, wyznaczonymi na podstawie części obserwowanej zasobu danych (próbki). Z próbki wyznaczano średnią wartość φ oraz ΔT i przyjmowano je jako imputowane φ i ΔT w przeciwpróbce węzłów cieplnych. Imputowany strumień przepływu w przeciwpróbce obliczano jako wynikowy z zależności (1) i (2).

Na rysunkach 1, 2 i 3 przedstawiono wyniki imputacji średnią dla przykładowej iteracji. Widoczne jest, że ten rodzaj imputacji powoduje w zbiorze danych imputowanych utratę zmienności parametrów przejawianą w rzeczywistości. Naturalnie najmniej trafna jest imputacja dla węzłów cieplnych, dla których w rzeczywistości parametry przyjmują wartości skrajne.



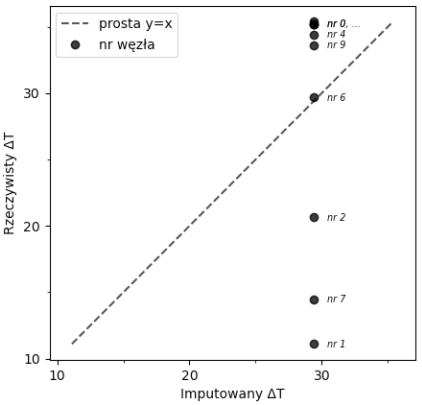
Rys. 1. Wartości φ faktyczne i imputowane w przykładowej iteracji

Fig. 1. Actual and imputed values of φ in an exemplary iteration



Rys. 2. Wartości strumienia przepływu wody faktyczne i imputowane w przykładowej iteracji

Fig. 2. Real and imputed values of volumetric flow of the water in an exemplary iteration



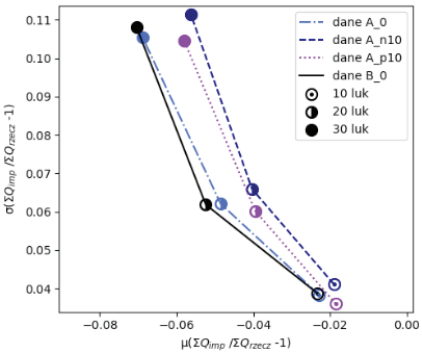
Rys. 3. Wartości ΔT faktyczne i imputowane w przykładowej iteracji

Fig. 3. Actual and imputed values of ΔT in an exemplary iteration

Kolejnym widocznym mankamentem metody jest to, że zmienne imputowane nie mogą przyjąć wartości wykraczających poza zakres wyznaczony przez najmniejszą i największą wartość obserwowaną w próbce.

Na rysunku 4 dokonano skrótowego przeglądu skuteczności imputacji mocy węzłów średnią. Wizualizowanym wskaźnikiem jest błąd względny wyznaczenia łącznej mocy węzłów cieplnych, stanowiący miarę trafności imputowania mocy w całej grupie węzłów. Na osi poziomej odwzorowano średnią wartość błędu względnego w 1000 wykonanych iteracjach, na osi pionowej odchylenie standardowe tego błędu. Pożądanymi cechami są: średnia błędu bliska zeru z jednoczesnym niskim odchyleniem standardowym, co oznaczałoby, że przeciętnie imputacja sumarycznej mocy węzłów jest trafna, a iteracje, w których implanty statystyczne znacznie odbiegają od wartości faktycznych, są rzadkie.

Imputacja, oceniana w ten sposób, najskuteczniejsza okazała się w wariantach z najmniejszą liczbą luk.



Rys. 4. Błąd względny wyznaczenia łącznej mocy węzłów cieplnych

Fig. 4. Relative error of district heating substations summary heat power

Dla każdego zestawu danych bazowych, wraz ze wzrostem liczby luk rośnie bezwzględna wartość średniego błędu oraz odchylenie standardowe błędu. Najmniejszy (co do wartości bezwzględnej) średni błąd, wynoszący –1,9%, otrzymano w zbiorze danych A_p10 dla 10 luk. Porównując wyniki między zbiorami danych bazowych, średni błąd był mniejszy w przypadku reprezentacji uśrednionych stanów systemu ciepłowniczego niż dla danych bazowych odwzorowujących stan chwilowy. Największy (co do wartości bezwzględnej) średni błąd, wynoszący –7%, otrzymano w zbiorze danych bazowych B_0, reprezentującym stan chwilowy, dla 30 luk, co jest przy 30 lukach wynikiem gorszym niż dla zbiorów danych bazowych odpowiadających stanom uśrednionym, dla których średni błąd był niższy.

Imputacja hot-deck

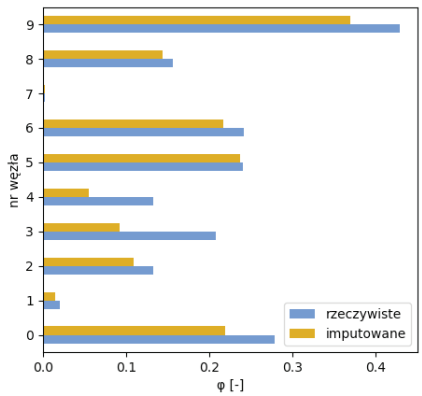
W ramach tej metody brakujące dane w przeciwpróbce uzupełniano wartościami pobranymi z najbardziej podobnych obiektów w próbce. Przetestowano dwa warianty metody. W pierwszym wariancie wyznacznikiem podobieństwa była moc zamówiona – za najlepszego donora dla danego węzła z przeciwpróbki uznawano węzeł z próbki o najbliższej mocy zamówionej. W drugim wariancie wyznacznikiem podobieństwa było zużycie ciepła przez dany węzeł w miesięcznym okresie. Oba warianty metody wymagają więc skorzystania z pomocniczego zasobu danych.

Motywacja stojąca za powyższymi dwoma wersjami metody jest następująca: nie zawsze moc zamówiona jest dobrana adekwatnie do faktycznego profilu zapotrzebowania węzła cieplnego. Zużycie ciepła w okresie miesięcznym jest wartością, która powinna być znana i łatwo dostępna jako dana pomocnicza nawet w przypadku braku godzinowych danych telemetrycznych, ponieważ na jej podstawie następuje rozliczenie z odbiorcą ciepła.

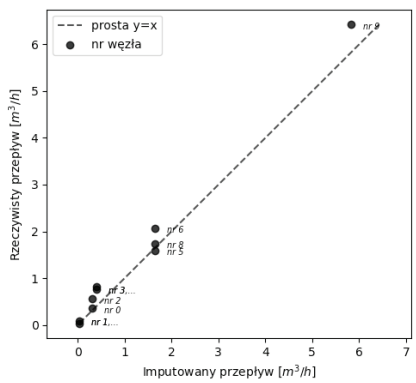
W obu wariantach od najbardziej podobnego donora pobierano wartość mocy, a następnie na podstawie w ten sposób zaimputowanej mocy wskazywano donora do pobrania wartości przepływu. Imputowany ΔT w przeciwpróbce obliczano jako wynikowy z zależności (1).

W każdym wariancie warunkowanym kombinacją liczby luk i zbioru danych bazowych skuteczniejsza okazywała się imputacja z wyznaczaniem donora opartym na zużyciu ciepła niż na mocy zamówionej. Wariant metody wykorzystującej jako miarę podobieństwa miesięczne zużycie ciepła przyjęto jako optymalną wersję metody hot deck do dalszych analiz.

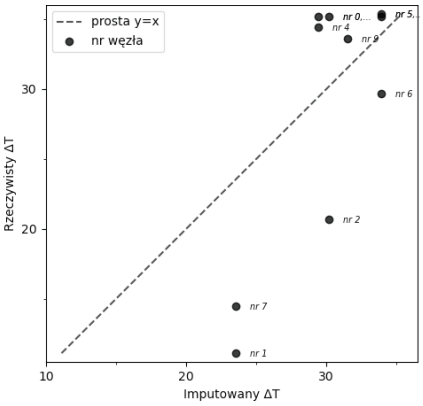
Na rysunkach 5, 6 i 7 przedstawiono wyniki imputacji hot-deck dla przykładowej iteracji. W porównaniu z imputacją średnią, zaletą metody hot-deck jest zachowanie w przeciwpróbce występującej w rzeczywistości zmienności parametrów. Tego typu imputacja ograniczona jest jednak do zamkniętej listy wartości występujących w próbce i, podobnie jak imputacja średnią, nie umożliwia przyjęcia wartości wykraczających poza granice najmniejszej i największej wartości obserwowanej w próbce.



Rys. 5. Wartości φ faktyczne i imputowane w przykładowej iteracji
Fig. 5. Actual and imputed values of φ in an exemplary iteration

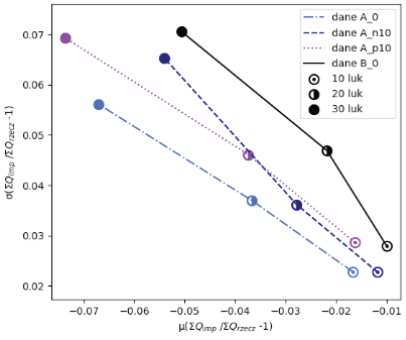


Rys. 6. Wartości strumienia przepływu wody faktyczne i imputowane w przykładowej iteracji
Fig. 6. Actual and imputed values of volumetric flow of the water in an exemplary iteration



Rys. 7. Wartości ΔT faktyczne i imputowane w przykładowej iteracji
Fig. 7. Actual and imputed values of ΔT in an exemplary iteration

Na rysunku 8 przedstawiono skrótowy przegląd skuteczności imputacji mocy metodą hot-deck, wizualizując średni w 1000 iteracjach błąd wyznaczenia sumarycznej mocy węzłów cieplnych i jego odchylenie standardowe.



Rys. 8. Błąd względny wyznaczenia łącznej mocy węzłów cieplnych
Fig. 8. Relative error of district heating substations summary heat power

Ponownie najtrafniejsze wyniki uzyskano dla wariantów z najmniejszym udziałem luk. Przy 10 lukach, średni błąd, w zależności od zbioru danych bazowych, waha się od –1% do –1,7%. W każdym zestawie danych bazowych, wraz ze zwiększeniem liczby luk, zwiększa się wartość bezwzględna błędu i jego odchylenia standardowego. Przeciętnie najbliższa rzeczywistości jest imputacja przeprowadzona na zbiorze danych reprezentującym chwilowy stan systemu ciepłowniczego, ale w porównaniu ze zbiorami danych opisującymi uśrednione stany systemu ciepłowniczego, zwykle większe jest odchylenie standardowe błędu, co należy interpretować jako większe ryzyko przeprowadzenia nietrafnej imputacji w przypadku niereprezentatywnej próbki. W zbiorze danych B_0 średni błąd wynosi –1%, –2,2% i –5,1% odpowiednio dla 10, 20 i 30 luk, tak więc wartość bezwzględna błędu jest dla każdej liczby luk mniejsza niż w innych zbiorach danych reprezentujących stany uśrednione systemu ciepłowniczego. Jednocześnie jednak w zbiorze B_0 dla każdej liczby luk wartości odchylenia standardowego są większe niż w innych zbiorach danych bazowych, sięgając poziomu 7,1% dla 30 luk, co jest najwyższą wartością otrzymaną w ramach metody hot-deck.

Imputacja regresją

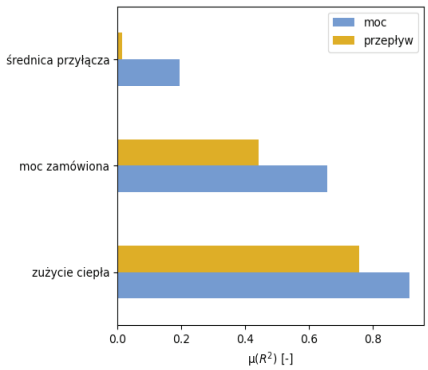
W ramach tej metody brakujące dane w przeciwpróbce uzupełniano w oparciu o model liniowej regresji wielorakiej wyznaczony na podstawie danych z próbki. Do próbki dopasowywano osobny model regresji liniowej opisujący moc węzłów cieplnych i osobny model opisujący strumień przepływu. Imputowany ΔT w przeciwpróbce obliczano jako wynikowy z zależności (1).

Zasadniczą kwestią jest decyzja, na jakich zmiennych powinny opierać się modele regresji liniowej. W tym celu przeprowadzono próbę, mającą na celu określenie, które zmienne mogą okazać się trafnymi predyktorami w ramach regresji. W 100 iteracjach sprawdzono trafność imputacji regresją opartą na jednej, dwóch lub trzech zmiennych. Wykorzystano katalog zmiennych z pomocniczych zasobów danych, możliwych do pozyskania w realistycznych scenariuszach:

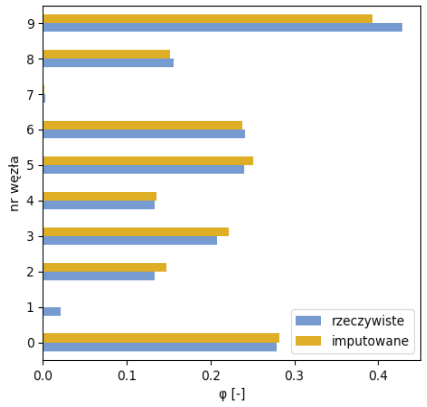
– podstawowe informacje charakteryzujące odbiorcę ciepła: moc zamówiona, miesięczne zużycie ciepła, liczba funkcji węzła cieplnego;
– informacje możliwe do pozyskania z baz danych GIS: liczba kondygnacji budynku, funkcja ogólna/dominująca/szczegółowa budynku;
– informacje możliwe do pozyskania z dokumentacji technicznej węzłów cieplnych: średnica przyłącza, model ciepłomierza, model zaworu regulacyjnego, model sterownika, model wymiennika ciepła c.o.

W zależności od charakteru zmiennych, były one uwzględniane w modelu regresji bezpośrednio (zmienne liczbowe) lub po przekodowaniu na wartości liczbowe (zmienne kategoryczne). Spośród powyższych zmiennych najlepszymi pojedynczymi predyktorami okazały się, jak pokazano na rysunku 9, miesięczne zużycie ciepła, moc zamówiona i średnica przyłącza.

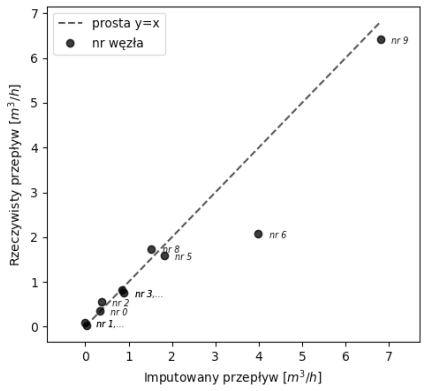
Najlepszą kombinacją zmiennych do imputacji mocy okazał się zbiór predyktorów złożony z miesięcznego zużycia ciepła i liczby kondygnacji budynku, dla którego w 100 testowych iteracjach średnie R^2 wyniosło 0,92.



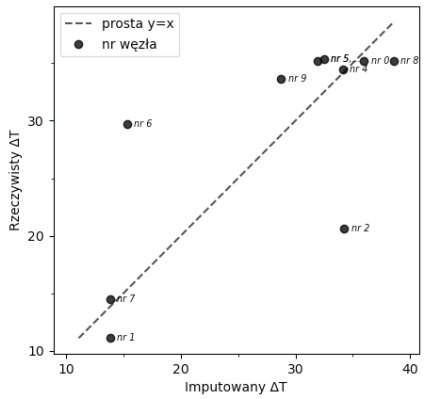
Rys. 9. Najlepsze pojedyncze predyktory wśród sprawdzonych zmiennych
Fig. 9. The best single predictors among the variables considered



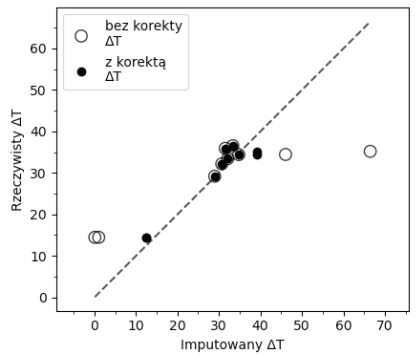
Rys. 10. Wartości φ faktyczne i imputowane w przykładowej iteracji
Fig. 10. Actual and imputed values of φ in an exemplary iteration



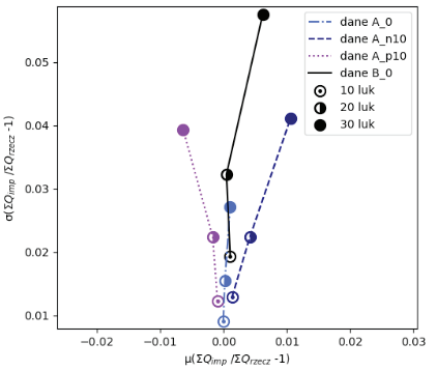
Rys. 11. Wartości przepływu faktyczne i imputowane w przykładowej iteracji
Fig. 11. Actual and imputed values of volumetric flow in an exemplary iteration



Rys. 12. Wartości ΔT faktyczne i imputowane w przykładowej iteracji
Fig. 12. Actual and imputed values of ΔT in an exemplary iteration



Rys. 13. Imputowane wartości ΔT z korektą i bez korekty w przykładowej iteracji
Fig. 13. Imputed values of ΔT with and without the correction in an exemplary iteration



Rys. 14. Błąd względny wyznaczenia łącznej mocy węzłów cieplnych
Fig. 14. Relative error of district heating substations summary heat power

przyjmując za imputowane ΔT odpowiednie wartości graniczne (maksymalną lub minimalną). Wykonując korektę ΔT, korygowano również imputowany przepływ, aby zachować zgodność tych parametrów z imputowaną mocą. Tego rodzaju korekta prowadziła do bardziej trafnej imputacji ΔT.

Najlepsze wyniki imputacji przepływu w testowych iteracjach osiągnięto dla kombinacji predyktorów: miesięczne zużycie ciepła i model sterownika; średnie R^2 wyniosło 0,96. Uzupełnienie powyższych predyktorów o dodatkową trzecią zmienną nie prowadziło do istotnej poprawy wyników. Powyżej przytoczone wartości R^2 zostały policzone w przeciwpróbce, stanowią więc miarę trafności imputacji, a nie dopasowania modelu regresji do próbki. (W testowych iteracjach występowały kombinacje zmiennych dające wyższe R^2 w próbce, nie przekładające się na lepsze R^2 w przeciwpróbce, co należy interpretować jako nadmiernie dopasowane, za mało ogólne modele regresji).

Na rysunkach 10, 11 i 12 przedstawiono wyniki imputacji regresją dla przykładowej iteracji. W porównaniu z wcześniej omawianymi metodami, imputowane wartości były przeciętnie najbardziej zgodne z wartościami faktycznymi. Zaletą tej metody jest zarówno występowanie w przeciwpróbce zmienności imputowanych wartości, jak i możliwość imputowania wartości niewystępujących w próbce, w tym wartości spoza zakresu wyznaczanego przez maksimum i minimum obecne w próbce.

Powyższa zaleta pociąga jednak za sobą pewną wadę. Mianowicie, może zostać imputowana kombinacja mocy i przepływu, która wynikowo prowadzi do wyznaczenia z zależności (1) niefizycznej wartości ΔT. Zobrazowano to na rysunku 13, gdzie widocznych jest kilka wynikowych wartości ΔT znacznie poniżej lub powyżej wartości spodziewanych dla analizowanego stanu systemu ciepłowniczego. Aby uniknąć niefizycznych wartości ΔT, posłużono się następującą metodą korekty. Do wartości ΔT w próbce dopasowywano rozkład Weibulla. Na dopasowanej gęstości tego rozkładu określano graniczne temperatury jako punkty wyznaczające dolne i górne 2,5% dystrybuanty. Jeśli imputowane ΔT wykraczały poza wyżej określone wartości graniczne, wykonywano korektę,

Na rysunku 14 przedstawiono skrótowo przegląd skuteczności imputacji mocy regresją, wizualizując średni w 1000 iteracjach błąd wyznaczenia sumarycznej mocy węzłów cieplnych i jego odchylenie standardowe. Ponownie najtrafniejsze wyniki uzyskano dla wariantów z najmniejszym udziałem luk. W każdym ze zbiorów danych bazowych najmniejszy co do wartości bezwzględnej błąd średni i najmniejsze odchylenie standardowe błędu otrzymano dla 10 luk. Najtrafniejsza była imputacja dla zbioru danych bazowych A_0, w którym średni błąd nie przekroczył 0,1% nawet dla 30 luk. Dla każdej liczby luk w zbiorze danych bazowych A_0 odchylenie standardowe błędu jest mniejsze niż w pozostałych zbiorach danych.

Porównanie metod imputacji

W tabeli 1 przedstawiono 5 wariantów prowadzących do najskuteczniejszej imputacji, ocenianej wedle wskaźnika ilustrującego trafność przewidywania sumarycznej mocy węzłów cieplnych w 1000 iteracjach. W tabeli 2 przedstawiono 5 wariantów prowadzących do najmniej skutecznej imputacji. Wartości liczbowe w obu tabelach odpowiadają skrajnym przypadkom zobrazowanym na rysunkach 4, 8 i 14.

Wyniki zostały posortowane według wartości średniej błędu względnego.

Najlepsze wyniki (najmniejszy co do wartości bezwzględnej średni błąd) otrzymano z wykorzystaniem metody regresji, dla liczby luk wynoszącej 10 lub 20. Największe co do wartości bezwzględnej błędy wystąpiły w przypadku liczby luk wynoszącej 30, przy wykorzystaniu imputacji średnią lub hot-deck. Najmniej trafne warianty imputacji, zestawione w tabeli 2, charakteryzują się większymi wartościami odchylenia standardowego błędu niż warianty najbardziej trafne przedstawione w tabeli 1, co należy interpretować jako większe ryzyko przeprowadzenia nietrafnej imputacji w przypadku niereprezentatywnej próbki.

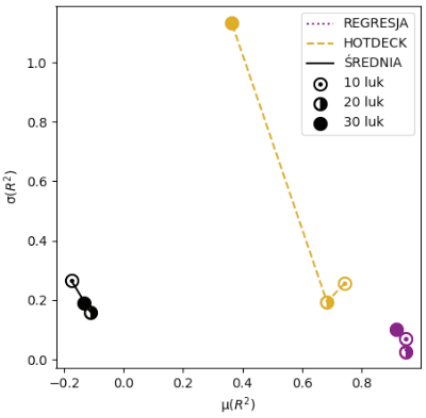
Tabela 1. Błąd względny wyznaczenia łącznej mocy węzłów cieplnych – 5 najbardziej trafnych wariantów imputacji
Table 1. Relative error of district heating substations summary heat power – 5 most accurate imputation scenarios

Lp.	Metoda	Dane bazowe	Liczba luk	Średni błąd względny [%]	Odchylenie standardowe błędu [%]
1	regresja	A_0	10	-0,01	0,9
2	regresja	A_0	20	0,02	1,5
3	regresja	B_0	20	0,05	3,2
4	regresja	A_p10	10	-0,1	1,2
5	regresja	A_n10	10	0,1	1,3

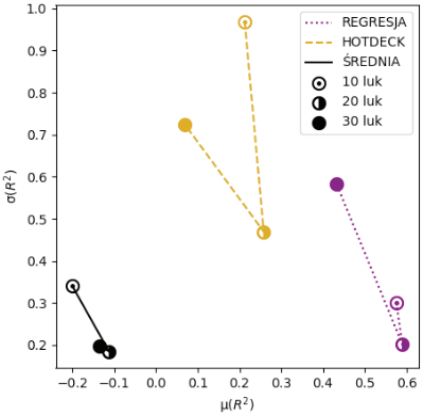
Tabela 2. Błąd względny wyznaczenia łącznej mocy węzłów cieplnych – 5 najmniej trafnych wariantów imputacji
Table 2. Relative error of district heating substations summary heat power – 5 least accurate imputation scenarios

Lp.	Metoda	Dane bazowe	Liczba luk	Średni błąd względny [%]	Odchylenie standardowe błędu [%]
1	hot-deck	A_p10	30	-7,4	6,9
2	średnia	B_0	30	-7,0	10,8
3	średnia	A_0	30	-6,9	10,5
4	hot-deck	A_0	30	-6,7	5,6
5	średnia	A_p10	30	-5,8	10,5

Na rysunkach 15 i 16 zwizualizowano trafność imputacji mocy różnymi metodami w 1000 iteracjach dla każdego z wariantów. Ze względu na rozstęp pomiędzy minimalnymi i maksymalnymi faktycznymi mocami, wskaźnik R^2 obliczono na podstawie wartości bezwymiarowej φ zdefiniowanej równaniem (2). Widoczne jest, że moc imputowana jest najtrafniej (największa średnia i najmniejsze odchylenie standardowe wskaźnika R^2) z wykorzystaniem regresji, zarówno w przypadku zbiorów danych reprezentujących stany systemu ciepłowniczego uśrednione, jak i chwilowe, przy czym lepiej wypada imputacja dla stanów uśrednionych – obserwowane są wyższe średnie wartości wskaźnika R^2 dla zbioru A_0 niż dla zbioru B_0.



Rys. 15. Wskaźnik R^2 imputacji φ w zbiorze danych bazowych A_0
Fig. 15. R^2 score of φ imputation in A_0 dataset



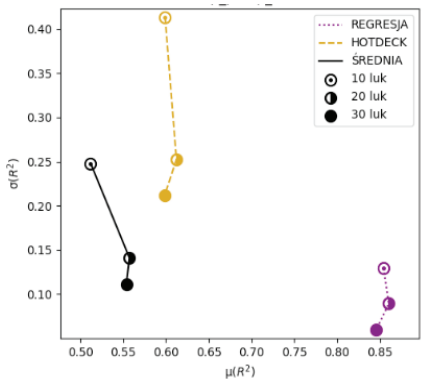
Rys. 16. Wskaźnik R^2 imputacji φ w zbiorze danych bazowych B_0
Fig. 16. R^2 score of φ imputation in B_0 dataset

Wśród rozważonych metod wyniki imputacji regresją najmniej zmieniają się wraz ze wzrostem udziału braków w danych. Dokładne wartości liczbowe przedstawiono w tabeli 3.

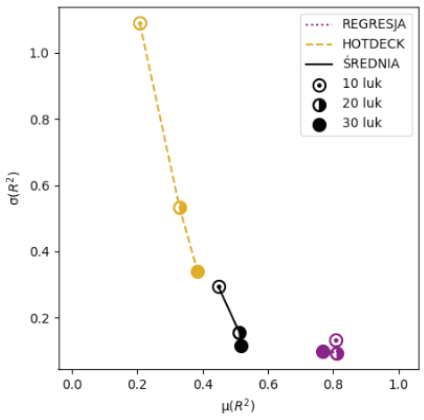
Tabela 3. Wskaźnik R^2 imputacji φ
Table 3. R^2 score of φ imputation

Metoda	Liczba luk	Średni wskaźnik R^2 imputacji φ [-]	
		Dane bazowe A_0	Dane bazowe B_0
regresja	10	0,95	0,57
regresja	20	0,95	0,59
regresja	30	0,92	0,43
hot-deck	10	0,74	0,21
hot-deck	20	0,68	0,26
hot-deck	30	0,36	0,07
średnia	10	-0,18	-0,20
średnia	20	-0,11	-0,11
średnia	30	-0,13	-0,13

Podobnie, jak przedstawiono na rysunkach 17 i 18, metoda regresji okazała się najskuteczniejsza (największa średnia i najmniejsze odchylenie standardowe wskaźnika R^2 w 1000 iteracjach) w imputacji przepływów. Dokładne wartości liczbowe przedstawiono w tabeli 4.



Rys. 17. Wskaźnik R^2 imputacji strumienia przepływu wody w zbiorze danych bazowych A_0
Fig. 17. R^2 score of water volumetric flow imputation in A_0 dataset



Rys. 18. Wskaźnik R^2 imputacji strumienia przepływu wody w zbiorze danych bazowych B_0
Fig. 18. R^2 score of water volumetric flow imputation in B_0 dataset

Tabela 4. Wskaźnik R^2 imputacji strumienia przepływu wody
Table 4. R^2 score of water volumetric flow imputation

Metoda	Liczba luk	Średni wskaźnik R^2 imputacji strumienia przepływu [-]	
		Dane bazowe A_0	Dane bazowe B_0
regresja	10	0,85	0,81
regresja	20	0,86	0,81
regresja	30	0,85	0,77
hot-deck	10	0,60	0,21
hot-deck	20	0,61	0,33
hot-deck	30	0,60	0,39
średnia	10	0,51	0,45
średnia	20	0,56	0,51
średnia	30	0,55	0,52

Najgorzej (najmniejsza średnia i największe odchylenie standardowe wskaźnika R^2 w 1000 iteracjach) wypadła imputacja dla ΔT, który w metodzie hot-deck i regresji jest zależny od trafności imputacji mocy i przepływu. Nawarstwiające się rozbieżności tych dwóch zmiennych względem wartości faktycznych kumulują się przy określaniu ΔT. Należy przy tym nadmienić, że choć imputacja ΔT była najmniej trafna i w przeciwpróbkach pojawiały się wartości znacznie odbiegające od faktycznych,

w przypadku przypisania do błędów określenia ΔT wag związanych z przepływami w węzłach, obserwowano znacznie lepszą trafność. Oznacza to trafniejszą imputację ΔT dla węzłów o większym wpływie na sieć ciepłowniczą.

Tabela 4. Wskaźnik R^2 imputacji ΔT – z błędami ważonymi strumieniem przepływu
Table 4. R^2 score of ΔT imputation – with errors weighted by volumetric flow

Metoda	Liczba luk	Średni wskaźnik R^2 imputacji ΔT [-]	
		Dane bazowe A_0	Dane bazowe B_0
regresja	10	0,97	0,91
regresja	20	0,98	0,93
regresja	30	0,98	0,92
hot-deck	10	0,91	0,83
hot-deck	20	0,86	0,79
hot-deck	30	0,78	0,73
średnia	10	0,64	0,62
średnia	20	0,69	0,68
średnia	30	0,69	0,69

Podsumowanie

Spośród rozważonych metod imputacji najlepiej sprawdziła się metoda oparta na regresji liniowej, osiągając wskaźnik dopasowania R^2 mocy węzłów cieplnych na poziomie do 0,95 i natężenia przepływu na poziomie do 0,86. Zaletą metody jest odzwierciedlenie w zbiorze imputowanym zmienności, która charakteryzuje dane faktyczne. Jednocześnie metoda ta umożliwia imputowanie wartości wykraczających poza katalog tych obserwowanych w próbce. Korzystne w przypadku imputacji regresją jest wykonywanie dodatkowej korekty ΔT, opartej na wartościach granicznych wyznaczonych z dopasowanego do próbki rozkładu Weibulla, aby uniknąć imputowania nierealistycznych wartości. Metoda imputacji regresją okazała się również, w porównaniu do pozostałych, stosunkowo odporna na wzrost udziału braków w danych. Stosunkowo niskie, w porównaniu z innymi metodami, wartości odchylenia standardowego analizowanych wskaźników, sugerują rzadkie występowanie mało trafnych iteracji, a zatem świadczą również o relatywnej odporności na niereprezentatywność próbek. Dla imputacji regresją odchylenie standardowe błędu względnego wyznaczenia sumarycznej mocy węzłów cieplnych wyniosło w grupie 5 najlepszych wariantów imputacji od 0,9% do 3,2%, podczas gdy dla 5 najgorszych wariantów, opierających się na innych metodach, odchylenie standardowe wyniosło od 5,6% do 10,8%.

Najlepszą grupą predyktorów dla imputacji przy pomocy regresji liniowej okazały się: dla mocy – miesięczne zużycie ciepła i liczba kondygnacji budynku (R^2 średnio na poziomie 0,92), a dla przepływów – miesięczne zużycie ciepła i model sterownika (R^2 średnio na poziomie 0,96). Jest to obserwacja ułatwiająca identyfikację potencjalnie cennej wiedzy pomocniczej (do uzupełniania braków w danych wejściowych modelu sieci ciepłowniczej) w innych realistycznie dostępnych zasobach danych, m.in. danych rozliczeniowych, bazach GIS i dokumentacji technicznej.

Metody imputacji średnią i hot deck sprawdziły się gorzej od imputacji opartej na regresji (wskaźnik dopasowania R^2 mocy węzłów na poziomie poniżej zera dla imputacji średnią i do 0,74 dla metody hot-deck oraz R^2 dla natężenia przepływu odpowiednio na poziomie do 0,56 i do 0,61), jednak w niektórych sytuacjach (uzupełnianie zbiorów danych uśrednionych, gdy występuje niewiele luk) metoda hot-deck może okazać się atrakcyjna, ze względu na łatwiejszą jej implementację niż imputacji opartej na regresji. W świetle wcześniej przedstawionych wyników, w przypadku metody hot-deck trafniejsze jest wyznaczanie adekwatnego „donora" w oparciu o miesięczne zużycie ciepła, nie o moc zamówioną.

Metody przedstawione w niniejszym artykule mogą stanowić przydatne narzędzie do uzyskania pełnego zbioru danych wejściowych dotyczących węzłów cieplnych, niezbędnych do wykonania obliczeń rozpływów wody i strat ciśnienia w sieci ciepłowniczej, w sytuacji braku godzinowych danych telemetrycznych z pewnej części odbiorów ciepła.

Przedstawione wyniki wskazują również na zgodną z intuicją zasadę, że trafniejsze są przewidywania oparte na pełniejszych danych. Podkreśla to istotność i wartość wnoszoną przez niezawodne i pokrywające możliwie dużą część odbiorców ciepła systemy telemetryczne.

Podziękowania

Praca została zrealizowana dzięki programowi Ministerstwa Nauki i Szkolnictwa Wyższego „Doktorat wdrożeniowy 2024", w ramach projektu DWD/0066/2024.

Składamy podziękowania firmie PGE Energia Ciepła S.A. za udostępnienie danych umożliwiających przeprowadzenie badań stanowiących podstawę niniejszego artykułu.

Literatura

[1] Niemyjski, O. Kalibracja modelu hydraulicznego sieci ciepłowniczej. INSTAL 2020, 415, 12-16, https://doi.org/10.36119/15.2020.3.1.
[2] Dhungana, H.; Bellotti, F.; Berta, R.; De Gloria, A. Performance Comparison of Imputation Methods in Building Energy Data Sets. In Applications in Electronics Pervading Industry, Environment and Society; Saponara, S., De Gloria, A., Eds.; Lecture Notes in Electrical Engineering; Springer International Publishing: Cham, 2021; Vol. 738, pp. 144–151 ISBN 978-3-030-66728-3.
[3] Pokropek, A. Wybrane statystyczne metody radzenia sobie z brakami danych. Polskie Forum Psychologiczne 2018, 291–310, https://doi.org/10.14656/PFP20180205.
[4] Młodak, A. Imputacja Danych w Spisach Powszechnych. WS 2010, 2010, 7–22, https://doi.org/10.59139/ws.2010.08.2.
[5] Enders, C.K.; Little, T.D. Applied Missing Data Analysis; Methodology in the social sciences; Second Edition.; The Guilford Press: New York London, 2022; ISBN 978-1-4625-4986-3.
[6] Wesołowski, J.; Tarczyński, J. Podstawy matematyczne technik imputacyjnych. Wiadomości Statystyczne 2016, 7–54.
[7] Misztal, M. Próba Oceny Wpływu Wybranych Metod Imputacji Danych Na Wyniki Klasyfikacji Obiektów z Wykorzystaniem Drzew Klasyfikacyjnych. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Taksonomia 2011, 18, 246--253.
[8] Horton, N.J.; Kleinman, K.P. Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician 2007, 61, 79–90, https://doi.org/10.1198/000313007X172556.
[9] Luo, Z.; Lin, X.; Zhong, W.; Feng, E. Research on Coarse Granularity Data Sample Completion Method for District Heating System. In Proceedings of the 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD); IEEE: Rio de Janeiro, Brazil, May 24 2023; pp. 249–254.
[10] Zhou, Y.; Aryal, S.; Bouadjenek, M.R. Review for Handling Missing Data with special missing mechanism, https://doi.org/10.48550/arXiv.2404.04905.
[11] Alwateer, M.; Atlam, E.-S.; El-Raouf, M.M.A.; Ghoneim, O.A.; Gad, I. Missing Data Imputation: A Comprehensive Review. JCC 2024, 12, 53–75, https://doi.org/10.4236/jcc.2024.1211004.
[12] Schaffer, M.; Tvedebrink, T.; Marszal Pomianowska, A. Three Years of Hourly Data from 3021 Smart Heat Meters Installed in Danish Residential Buildings. Sci Data 2022, 9, 420, https://doi.org/10.1038/s41597-022-01502-3.
[13] Johra, H.; Leiria, D.; Heiselberg, P.; Marszal-Pomianowska, A.; Tvedebrink, T. Treatment and Analysis of Smart Energy Meter Data from a Cluster of Buildings Connected to District Heating: A Danish Case. E3S Web Conf. 2020, 172, 12004, https://doi.org/10.1051/e3sconf/202017212004.
[14] Leiria, D.; Johra, H.; Marszal-Pomianowska, A.; Pomianowski, M.Z.; Kvols Heiselberg, P. Using Data from Smart Energy Meters to Gain Knowledge about Households Connected to the District Heating Network: A Danish Case. Smart Energy 2021, 3, 100035, https://doi.org/10.1016/j.segy.2021.100035.
[15] Inman, D.; Elmore, R.; Bush, B. A Case Study to Examine the Imputation of Missing Data to Improve Clustering Analysis of Building Electrical Demand. Building Services Engineering Research and Technology 2015, 36, 628–637, https://doi.org/10.1177/0143624415573215.
[16] Lee, G.; Choi, S.; Choi, Y.; Koo, J.; Kim, D.-W.; Yoon, S. A Two-Stage Imputation Method for Enhancing Urban Building Energy Data Resilience Using Bayesian Inference. Energy and Buildings 2025, 349, 116515, https://doi.org/10.1016/j.enbuild.2025.116515.