POLITECHNIKA WARSZAWSKA

Wydział Mechaniczny
Energetyki i Lotnictwa

ROZPRAWA DOKTORSKA

mgr inż. Teresa Kurek

System generujący prognozy zapotrzebowania na ciepło dla Warszawskiej Sieci Ciepłowniczej

Promotor
prof. dr hab. inż. Konrad Świrski

Promotor pomocniczy
dr inż. Konrad Wojdan

WARSZAWA 2021

Streszczenie

W pracy zaprezentowano podstawy teoretyczne oraz realizację praktyczną systemu SPROG wyznaczającego prognozy zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej. Algorytmy wyznaczające wartości prognoz zostały stworzone na podstawie modeli uczenia maszynowego wykorzystujących pomiary konsumpcji ciepła przez użytkowników sieci oraz dane pogodowe dla Warszawy. Wykorzystane dane pomiarowe zawierały pomiary z około 16 000 węzłów cieplnych rejestrowanych z rozdzielczością godzinową. W niniejszej rozprawie przedstawiono metody prognozowania szeregów czasowych oraz metryki pozwalające na ocenę jakości stworzonych modeli predykcyjnych. Opisano proces tworzenia systemu bazującego na technikach uczenia maszynowego składającego się z trzech etapów: definicji problemu, modelowania oraz implementacji. W pracy zaprezentowano poszczególne warstwy systemu oraz ich kluczowe algorytmy pozwalające na walidację danych pomiarowych, skalowanie brakujących pomiarów, wyznaczanie zapotrzebowania na ciepło dla całej sieci lub jej zdefiniowanych obszarów oraz trenowanie i ewaluację modeli prognostycznych. Finalny model prognostyczny zapotrzebowania na ciepło dla całej sieci stanowi 4-warstwowa sztuczna sieć neuronowa z autoregresyjnym wejściem, natomiast wynik prognozy zawiera szacowaną wartość zapotrzebowania na ciepło wraz z wartościami przedziałów ufności prognozy. Jakość modeli prognostycznych oceniano indywidualnie dla trzech typowych sezonów, tj. sezonu letniego, sezonu grzewczego oraz sezonu przejściowego. W pracy wykazano, iż możliwe jest uzyskanie precyzyjnych prognoz dla każdego z analizowanych sezonów. Przedstawiona w niniejszej rozprawie analiza wyników wdrożenia systemu w warszawskiej sieci ciepłowniczej potwierdza skuteczność prezentowanego rozwiązania.

Słowa kluczowe: *sieci ciepłownicze, prognoza zapotrzebowania na ciepło, sztuczne sieci neuronowe*

Abstract

The system that generates the heat demand forecasts for the Warsaw Ddistrict Heating Network

This thesis presents theoretical background and practical implementation of the system which generates the heat demand forecasts for the Warsaw District Heating Network. The algorithms that evaluate the forecasts have been developed based on machine learning models where the input data are the measurements of heat consumption by the end users and the weather data. The dataset contains the one hour resolution measured values from 16 000 substations. This thesis presents the methods for time series forecasts and the metrics allowing the evaluation of the prediction models. The process of constructing the machine learning based system has been described and it consists of three main stages: problem definition, modelling and model implementation. This thesis presents each system layer and its key algorithms allowing the efficient data workflow, that is: data validation, missing data replacement, estimation of heat demand for the whole grid or its areas as well as model training and evaluation. The final forecast model has a form of the artificial neural network with 4 hidden layers and autoregressive input, whereas the forecast consists of the forecast value and the confidence intervals. The models quality has been evaluated individually for three distinctive seasons of the year, i.e. heating, summer and transient season. The thesis proves that it is possible to precisely forecast heat demand for each analyzed season. The analysis of the outcomes from the implementation in the Warsaw District Heating Network demonstrates the efficiency and effectiveness of the solution.

Keywords: *district heating networks, heat demand forecast, artificial neural networks*

Spis treści

Rozdział 1

Wprowadzenie

Systemy Ciepłownicze (SC) są jednym z najlepszych narzędzi efektywnego wytwarzania oraz dystrybucji ciepła w dużych i gęsto zaludnionych obszarach miejskich. Według raportu Heat Roadmap [56], oceniającego ekonomiczną przydatność sieci ciepłowniczych w Unii Europejskiej, 78 proc. zapotrzebowania na ciepło w budynkach pochodzi z gęsto lub bardzo gęsto zaludnionych obszarów miejskich, na których mieszka 70 proc. populacji. Raport określa obecny udział miejskich sieci ciepłowniczych w rynku ciepła na poziomie 12 proc., jednocześnie estymując że z ekonomicznego punktu widzenia optymalny poziom wynosi 50 proc.. Wspomniany raport stanowi uzupełnienie raportu Heat Roadmap (HRE4) [50], który określa sieci ciepłownicze jako kluczowy czynnik mający potencjalny wpływ na transformację w kierunku gospodarki niskoemisyjnej stanowiącej podstawę wydajnych systemów energetycznych, natomiast jako pierwszy krok w realizacji transformacji wskazano konieczność skoncentrowania się na oszczędzaniu ciepła oraz zwiększaniu udziału miejskich sieci ciepłowniczych. Oszczędności ciepła w systemach ciepłowniczych możliwe są do uzyskania w wyniku efektywniejszej dystrybucji ciepła lub poprawy efektywności energetycznej odbiorców ciepła. Poprawa efektywności energetycznej odbiorców wiąże się z perspektywą budowy budynków niskoenergetycznych, która została przeanalizowana i opisana w licznych pracach badawczych [73, 41], zawierając koncept budynków energooszczędnych [4] oraz budynków o zerowej emisji i domy plus energetyczne [62, 47]. Jednakże prezentowane rozwiązania są przeznaczone głównie do nowo powstających budynków, a nie istniejących już obiektów, które ze względu na długą żywotność będą stanowić główną część zapotrzebowania na ciepło przez wiele nadchodzących dziesięcioleci. Literatura naukowa porusza kwestię zmniejszenia zapotrzebowania na ciepło w istniejących budynkach, konkludując że taki wysiłek

wiąże się ze znacznymi kosztami inwestycyjnymi [80], natomiast nie przeprowadzono badań, które wskazywałyby, jak całkowicie wyeliminować zapotrzebowanie na ciepło w istniejących budynkach w racjonalnie określonych ramach czasowych.

Scenariusze transformacji sieci ciepłowniczych opracowane w HRE4 opierają się na wyborach projektowych bazujących na wykorzystaniu sprawdzonych oraz już istniejących technologii w opozycji do radykalnych działań lub innowacyjnych, zaplanowanych projektach technologicznych. Idea HRE4 polegała na zaprojektowaniu systemu, który pokaże, w jaki sposób można stworzyć niskoemisyjną gospodarkę przy użyciu istniejących technologii, które nie stanowią bariery w ewentualnym przejściu na system bazujący wyłącznie na energii odnawialnej. W przypadku ciepłownictwa oznacza to, że nowe technologie dotyczą głównie ciepłownictwa trzeciej generacji (3GDH) z kierunkiem ku wprowadzeniu ciepłownictwa IV generacji (4GDH).

Trend zmian na przestrzeni trzech generacji systemów ciepłowniczych kierowany był w stronę niskotemperaturowej dystrybucji ciepła uzyskiwanej dzięki modernizacji rurociągów przesyłowych. By móc spełnić swoją rolę w przyszłych zrównoważonych systemach energetycznych, ciepłownictwo IV generacji będzie musiało sprostać następującym wyzwaniom [46]:

1. Zdolność do dostarczania niskotemperaturowego ciepła do ogrzewania pomieszczeń i ciepłej wody użytkowej do istniejących budynków, istniejących budynków poddanych renowacji energetycznej oraz nowych budynków niskoenergetycznych.

2. Zdolność do dystrybucji ciepła w sieciach o małych stratach sieciowych.

3. Zdolność do odzysku ciepła ze źródeł niskotemperaturowych i integracji odnawialnych źródeł ciepła, takich jak kolektory i źródła geotermalne.

4. Zdolność do bycia integralną częścią inteligentnych systemów energetycznych (tj. zintegrowanych inteligentnych sieci elektrycznych, gazowych i cieplnych), włączając w to bycie integralną częścią systemów sieci chłodniczej czwartej generacji.

5. Umiejętność zapewnienia odpowiednich struktur planistycznych, kosztowych i motywacyjnych w działalności oraz strategicznych inwestycji związanych z transformacją w przyszłe zrównoważone systemy energetyczne.

W trzeciej generacji parametry sieci dopasowane były do źródła ciepła, do jego sprawności i możliwych do osiągnięcia temperaturach czynnika. W przypadku czwartej generacji projektowanie systemu zaczyna się od odbiorcy. Jednym z kroków zmierzających do przekształcenia sieci ciepłowniczych jest wprowadzenie telemetrii pozwalającej na bieżące odczytywanie i przesyłanie pomiarów z opomiarowanych obiektów sieci ciepłowniczych. Bieżące odczytywanie, przesyłanie oraz przechowywanie danych dotyczących zużycia ciepła w węzłach cieplnych danej sieci stanowi podstawę do szacowania zapotrzebowania na ciepło w całej sieci oraz jej obszarach. Na podstawie zapotrzebowania odbiorcy i charakteru odbioru ciepła szacuje się straty przesyłu sieci i wynikowo uzyskiwane jest zapotrzebowanie na moc w źródle ciepła. W związku z powyższym precyzyjna średnio- i krótkoterminowa prognoza zapotrzebowania na ciepło jest jednym z kluczowych elementów koniecznych do poprawy planowania i eksploatacji systemu ciepłowniczego IV generacji [46]. Prognoza zapotrzebowania na ciepło użyteczna z punktu widzenia 4GDH musi spełniać następujące warunki:

- prognoza powinna obejmować modele różnych odbiorców ciepła (całe sieci, obszary sieci oraz indywidualni odbiorcy) aby umożliwić szeroki zakres zastosowań.
- prognoza powinna być łatwa do wdrożenia i zarządzania przez systemy wspierające operatorów i menadżerów sieci ciepłowniczych.
- prognoza musi być dostosowywana do wahań sezonowych, a także do zmieniających się zachowań konsumentów.

Podsumowując, metoda prognozowania krótkoterminowego obciążenia cieplnego musi być ogólnie stosowalna, prosta i adaptacyjna.

Temat prognzowana zapotrzebowania na ciepło, badawczo nie jest tematem nowym. Pierwsze prace przedstawiające modele prognostyczne pojawiły się w latach 90-ych XX wieku [74] i na przestrzeni lat liczba prezentowanych modeli sukcesywnie rosła, obejmując szeroki zakres klas, zarówno modeli liniowych [9] jak i modeli nieliniowych [24, 31]. Wraz z rozwojem technologii dostępność narzędzi obliczeniowych badana była również w bardziej skomplikowany sposób, a jednocześnie wykorzystywała obliczeniowo trudniejsze modele [76]. Wyniki modeli odnoszą się w znaczniej mierze do małego wycinka roku, np. wybranych dni lub tygodni, dodatkowo badania przedstawiane w literaturze naukowej skupiają się na pełnym sezonie grzewczym.

Wykorzystanie prognozy zapotrzebowania przez przedsiębiorstwo zajmujące się dystrybucją ciepła wymaga kompleksowego rozwiązania, które nie ogranicza się wyłącznie do stworzenia modelu prognostycznego, w szczególności gdy prognoza ma za zadanie wspomagać codzienną pracę operatorów lub być elementem algorytmu optymalizacji pracy sieci. Tworzenie systemów pozwalających na bieżące generowanie prognozy możliwe jest dzięki dynamicznemu rozwojowi uczenia maszynowego, sztucznej inteligencji, telemetrii oraz mocy obliczeniowych komputerów. Możliwości realnego wykorzystania prognozy zapotrzebowania na ciepło w inteligentnych sieciach ciepłowniczych nie zostały obszernie przebadane i dostępne są nieliczne prace prezentujące to zagadnienie [51].

W niniejszej rozprawie autorka przedstawia system generujący prognozy zapotrzebowania na ciepło (SPROG) stworzony we współpracy z firmą Transition Technologies S.A., Transition Technologies Science S.A. oraz Veolia Energia Warszawa S.A. który stanowi autonomiczną warstwę systemu wsparcia decyzji (SWD). SWD stanowi kluczową część projektu Inteligentna Sieć Ciepłownicza (ISC) realizowanego przez Veolia Eneria Warszawa S.A. Podstawowym założeniem projektu było optymalizowanie pod względem kosztowym procesu dystrybucji ciepła w Warszawie oraz obniżenie emisji CO_2, założenia te zostały zrealizowana z wykorzystanie SWD. Na podstawie bieżących oraz historycznych danych pomiarowych z elementów sieci ciepłowniczych, takich jak wymienniki ciepła czy przepompownie, algorytmy SWD przewidują zapotrzebowanie mieszkańców na moc cieplną oraz proponują dyspozytorowi optymalny scenariusz sterowania siecią, który pozwoli na dostawę ciepła zgodnie z oczekiwaniami mieszkańców przy jednoczesnej minimalizacji jego strat. Sam system prognostyczny generuje prognoze zapotrzebowania na ciepło dla całej sieci, która stanowi jedną z danych wejściowcyh do algorytmu optymalizacji jej pracy oraz prognozy zapotrzebowania na ciepło dla obszarów sieci i indywidualnych odbiorców.

Przedstawione powyżej zadania SPROG pozwalają na realizację zadań będących składnikami inteligentnego systemu ciepłowniczego:

- optymalizacja pracy sieci;
- opracowywanie różnych scenariuszy pracy sieci, przedziałów ufności pozwalających na określenie zakresu wartości w jakim uplasuje się rzeczywiste zapotrzebowanie z 90 proc. prawdopodobieństwem. Przykładowo, wykorzystując w zadaniu optymalizacji górną granicę przedziału ufności prognozy wyznaczone wartości pracy sieci będą wartościami bezpiecznymi pozwalającymi na dostawę ciepła bliską pewności jednakże straty ciepła

w takim przypadku są znaczące. Z drugiej strony dolna wartość przedziału ufności może określać minimalną ilość ciepła, jaką należy dostarczać do sieci taki scenariusz będzie scenariuszem ekonomicznym (bardzo niskie straty ciepła, brak zamawiania nadwyżki ciepła) jednak nie zapewnia z dużą dozą pewności odpowiedniej ilości ciepła wszystkim odbiorcom;

- efektywniejsze zarządzanie awariami pojawiającymi się w sieci – zarówno lokalnie, jak i w większej skali.

Wyniki uzyskane w trakcie tworzenia systemu [43], a następnie potwierdzone po wdrożeniu, pokazują że stworzone modele osiągają dokładność podobną lub lepsza lepszą z modelami stworzonymi dla innych sieci ciepłowniczych. SPROG przez implementację w rzeczywistym systemie stanowi kolejny krok poprawiania pracy sieci ciepłowniczych trzeciej generacji oraz możliwości wdrożenia sieci ciepłowniczych czwartej generacji.

1.1 Cel i zakres pracy

Celem badawczym pracy było opracowanie systemu generującego prognozy zapotrzebowania na ciepło (SPROG) dla warszawskiej sieci ciepłowniczej (WSC). SPROG stanowi autonomiczną warstwę systemu wsparcia decyzji (SWD), którego głównym celem jest optymalizacja pracy sieci ciepłowniczej. Generowane prognozy stanowią informację wejściową do optymalizatora pracy sieci ciepłowniczej. SPROG spełnia następujące założenie główne oraz założenia dodatkowe systemu.

Założenie główne systemu:

- System generuje prognozy zapotrzebownia na ciepło dla całej sieci.
- System generuje prognozę zapotrzebowania na ciepło dla zadanego horyzontu czasowego wraz z przedziałami ufności.

Założenia poboczne systemu:

- System generuje prognozy zapotrzebowania na ciepło wraz z przedziałami ufności dla obszarów sieci.

Zarówno założenia główne, jak i założenia poboczne systemu wymagają stworzenia suplementarnych warstw zapewniających realizację jego założeń. Wymagane jest stworzenie następujących warstw:

1. warstwa wczytywania danych,
2. warstwa walidacji oraz uzupełniania brakujacych danych,
3. warstwa agregacji danych,
4. warstwa wyznaczająca prognozy.

1.2 Tezy pracy

Mając na uwadze przedstawiony w rozdziale 1.1 cel oraz zakres pracy sformułowano następujące tezy:

Teza główna 1.2.1. *Możliwe i konieczne jest stworzenie systemu realizującego prognozy zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej na podstawie danych pomiarowych z systemu telemetrii oraz danych pogodowych.*

Teza poboczna 1.2.2. *Możliwe i wskazane jest prognozowanie zapotrzebowania na ciepło dla obszarów sieci wykorzystując telemetryczne pomiary z węzłów cieplnych.*

Teza poboczna 1.2.3. *Możliwe jest uzyskanie wysokiej dokładności prognoz zapotrzebowania na ciepło dla dużych sieci ciepłowniczych dzięki tworzeniu modeli prognostycznych bazujących na metodach uczenia maszynowego. Wysoka dokładność prognoz możliwa jest do uzyskania dla każdego okresu w roku.*

Udowodnienie przedstawionej tezy głównej oraz tez pobocznych stanowi cel rozprawy.

1.3 Układ rozprawy

W rozdziale drugim przedstawiono wybrane metody prognozowania szeregów czasowych. Metody te wykorzystane zostały w procesie doboru oraz trenowania modeli prognostycznych

będących podstawą systemu SPROG. W kolejnej części rozdziału przedstawiono wskaźniki jakości prognoz, za pomocą których możliwa jest ewaluacja tworzonych modeli oraz porównanie uzyskanych wyników z wynikami prezentowanymi w literaturze naukowej.

Aspekty prognozowania specyficzne dla modelowania zapotrzebowania na ciepło przedstawiono w pierwszej części rozdziału trzeciego. W drugiej części rozdziału przedstawiono przegląd literatury dotyczący prognozowania zapotrzebowania na ciepło dla sieci ciepłowniczych, który podzielony został wedle wspomnianych aspektów prognozowania. Przedstawiona analiza pozwala ocenić i określić wymagany do uzyskania przez system SPROG poziom referencyjny dokładności modeli prognostycznych.

Dlatego że SPROG stworzony został, bazując na danych telemetrycznych w rozdziale czwartym przedstawiono metodologię budowy systemów bazujących na technikach uczenia maszynowego.

W rozdziale piątym przedstawiono podstawowe informacje dotyczące warszawskiej sieci ciepłowniczej.

W rozdziale szóstym przedstawiono poszczególne etapy tworzenia systemu SPROG, zgodnie z metodologią przedstawioną w rozdziale czwartym. Przedstawiono wyniki modeli prognostycznych dla zbiorów danych testowych oraz porównano je z wynikami prezentowanymi w literaturze..

W rozdziale siódmym przeanalizowano wyniki wdrożeń SPROG w warszawskiej sieci ciepłowniczej. Wyniki te potwierdzają skuteczność przedstawionego w niniejszej pracy rozwiązania.

W ostatnim rozdziale podsumowano dotychczasowe rozwiązania.

Rozdział 2

Prognozowanie szeregów czasowych

Prognozowanie (predykcja) jest racjonalnym i naukowym przewidywaniem przyszłych zdarzeń. W zależności od dostępnych danych empirycznych oraz natury badanego zjawiska, np.: popytu na gotowe wyroby, wielkości sprzedaży i produkcji oraz wielkości zapasów, istotny jest dobór adekwatnej metody prognozowania. Metody prognozowania to specjalnie określone metody postępowania wykorzystywane do rozwiązywania zadań prognostycznych. Równocześnie to sposób przetwarzania danych implikuje regułę wyznaczania prognozy. Wobec tego metoda prognozowania składa się z dwóch elementów: modelu oraz reguły prognozowania [68].

Istnieje wiele podziałów prognoz na klasy w zależności od przyjętego kryterium podziału. Najczęściej stosowane podziały prognoz to klasyfikacja z uwagi na:

- typ zmiennej prognozowanej:
  - metody ilościowe (punktowe, przedziałowe), dotyczą zmiennych mierzalnych wyrażanych wartością liczbową. Prognoza ilościowa może być punktowa lub przedziałowa;
  - metoda jakościowa, dotyczy zmiennej niemierzalnej formułowanej przy pomocy opisu;
- horyzont prognozy:
  - prognoza krótkookresowa;
  - prognoza średniookresowa;
  - prognoza długookresowa;
- cel prognozy:

- prognoza poznawcza – służy rozwojowi teorii;
- prognoza informacyjna – stanowi podstawę procesów planistycznych;
- prognoza ostrzegawcza – informuje o niekorzystnym przebiegu badanych zjawisk;
- prognoza normatywna –prezentuje pożądane stany procesów.

Najważniejszym celem prognozowania jest wspomaganie procesów decyzyjnych w przedsiębiorstwach, dlatego wyróżnia się podstawowe i pomocnicze funkcje prognoz. Do podstawowych funkcji prognoz zalicza się:

- funkcję preparacyjną, która jest działaniem przygotowującym do podejmowania kolejnych kroków;
- funkcję aktywizującą, która pobudza do podejmownia działań sprzyjających realiazacji korzystnej prognozy lub przeciwstawia się spełnieniu negatywnej prognozy;
- funkcję informacyjną, która przyzwyczaja społeczeństwo do nadchodzących zmian.

Natomiast do pomocniczych funkcji prognoz zalicza się:

- funkcję argumentacyjną, która polega na dostarczaniu przez prognozę argumentów do podejmowania decyzji;
- funkcję doradczą, która przygotowuje odpowiednie informacje odnoszące się do badanych zjawisk będących składową procesu decyzyjnego;
- fuckcję mediacyjną, która ukazuje pomocność prognozy.

Sam proces prognozowania składa się z dwóch zasadniczych faz: diagnozowania przeszłości oraz określania przyszłości. Faza diagnozowania przeszłości ma na celu poznanie natury badanego zjawiska oraz mechanizmów jego funkcjonowania w tym identyfikację kształtujących je czynników. Określenie przyszłości to przejście od analizy danych oraz budowy modelu do prognozy wyznaczanej zgodnie ze zdefiniowaną regułą prognozowania.

W procesie prognozowania kluczową rolę odgrywają dane pomiarowe, są one podstawą wyboru klasy modelu prognostycznego, który opisuje i wyjaśnia zależności między zmienną prognozowaną oraz zmiennymi niezależnymi. By stworzyć wiarygodny model prognostyczny, dostępne dane statystyczne muszą posiadać określone właściwości, takie jak: jednorodność, kompletność, aktualność dla przyszłości oraz wiarygodność [68].

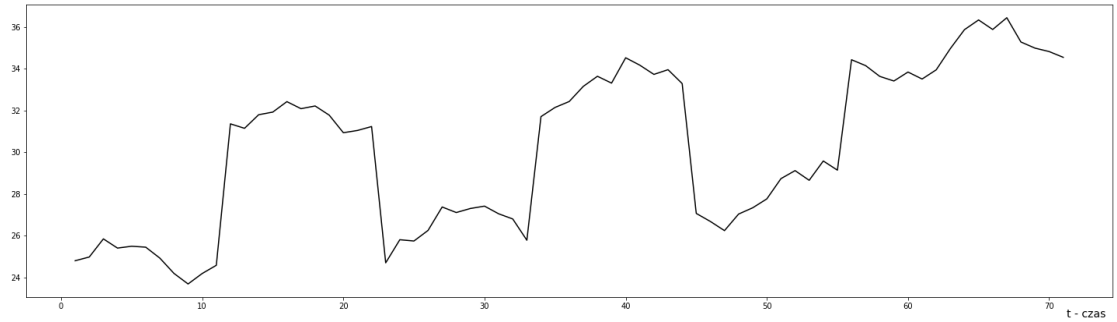
Prognozy realizowane w praktyce gospodarczej najczęściej odnoszą się do zjawisk dynamicznych oraz ciągłych w czasie, które reprezentowane są przez zbiór danych będących szeregiem czasowymm tj. zbiór mierzonych wartości, uporządkowanych chronologicznie w jednakowych, następujących po sobie przedziałach czasu. Realizacja prognoz dla szeregów czasowych wymaga znajomości teori szeregów czasowych oraz odpowiednich modeli prognostyczmych.

2.1 Teoria szeregów czasowych

Teoria szeregów czasowych daje możliwość modelowania oraz prognozowania różnych procesów [57] i jest obecnie powszechnie stosowana w wielu dziedzinach, takich jak inżynieria [6, 15, 54, 64, 66, 53], medycyna [16, 63, 69], bioinformatyka [34], ekonomia [17, 48, 7, 71]. Szereg czasowy jest to ciąg obserwacji uporządkowanych w czasie, których pomiary wykonywane są z dokładnym krokiem czasowym. Oznaczając przez $t(t-1,...,n)$ momenty czasu w których obserwowano wartość zmiennej, a przez y_t wynik obserwacji, szereg czasowy zapisujemy jako zbiór:

$$\{y_t; t = 1...,n\} \tag{2.1}$$

Przykładowy szereg czasowy jest przedstawiony na wykresie 2.1.



Rysunek 2.1: Przykładowy szereg czasowy.

W szeregu czasowym wyróżniamy dwie główne składowe:

- składowa systematyczna, która jest możliwa do objaśnienia w oparciu o system, w którym budowany jest model. Elementy składowej systematycznej to:

  - trend (tendencja rozwojowa) wyrażająca długookresową skłonność do jednokierunkowych zmian (wzrostu lub spadku) wartości badanej zmiennej;
  - składowa stała wyrażająca stały, niezmienny poziom wartości badanego zjawiska;
  - wahania sezonowe i cykliczne obrazujące cykliczne wahania wartości szeregu wokół tendencji rozwojowej. Wahania sezonowe są wahaniami o okresie nieprzekraczającym jednego roku i reprezentują pewne efekty powtarzające się z pewną prawidłowością, co roku w tych samych okresach;

- składowa przypadkowa (szum, wahania przypadkowe) zawierająca przypadkowe wahania szeregu czasowego wokół części systematycznej, które trudno zidentyfikować.

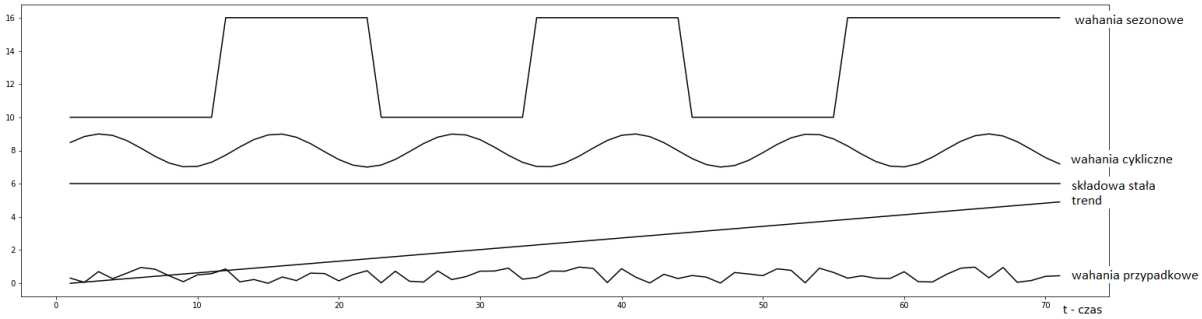
Analizując szereg czasowy dokonuje się dekompozycji szeregu czasowego, czyli wyodrębnienia poszczególnych składowych danego szeregu. Często pierwszą dekompozycję szeregu czasowego dokonuje się na podstawie obserwacji wzrokowej. Graficzne przedstawienie dekompozycji przykładowego szeregu czasowego przedstawiono na wykresie 2.1, a przedstawienie jego dekompozycji zaprezentowano na wykresie 2.2.

Dokonując analizy i dekompozycji szeregu czasowego należy pamiętać, że nie zawsze występują wszystkie składowe szeregu. Ponadto zdarzają się także obserwacje nietypowe, które mogą wynikać z wielu czynników takich jak: znaczne wahania losowe, błąd w rejestracji danych, gwałtowna zmiana badanego zjawiska. Marginalnym, jednakże spotykanym przypadkiem jest zmiana kierunku tendencji rozwoju szeregu czasowego, w takim przypadku wymagane jest zastosowanie specjalnych metod prognozowania. Takie szeregi czasem okazują się niemożliwe do prognozowania.

Szczególne znaczenie w kontekście jakości uzyskiwanych wyników ma identyfikacja modelu. Model powinien możliwie najlepiej opisywać naturę badanego zjawiska, dlatego w procesie tworzenia modelu konieczna jest weryfikacja różnych hipotez oraz modeli.

2.2 Modele szeregów czasowych

Prognozowanie szeregów czasowych wymaga przetworzenia informacji o przeszłości oraz zbudowania na postawie uzyskanych wyników odpowiedniego modelu formalnego. Modelem szeregu czasowego służącym do prognozowania przyszłej wartości zmiennej prognozowanej y w momencie prognozowania t, tj. $\widehat{y_t}$ jest model formalny, którego zmiennymi wejściowymi są



Rysunek 2.2: Dekompozycja przykładowego szeregu czasowego.

zmienna czasowa oraz przeszłe wartości lub prognozy zmiennej y_t. Dodatkowymi wejściami modelu mogą być aktualne oraz przeszłe wartości zmiennych niezależnych $x1,...,xn$. Ogólna postać modelu może być następująca:

$$\begin{aligned}\widehat{y_t} = f(y, y_{t-1}, ..., y_{t-p}, y_{t-1}, ..., y_{t-p}, y^*_{t-1}, ..., \widehat{y_{t-p}}, \\ x1_{t-1}, ..., x1_{t-p}, ..., xn_{t-1}, ..., xn_{t-p}, \zeta_t)\end{aligned} \tag{2.2}$$

Obecnie w praktyce prognozowania szeregów czasowych bazującej na technikach uczenia maszynowego stosuje się kilka wiodących metod. Modele liniowe są podstawowymi i powszechnie stosowanymi typami modeli predykcyjnych, nie uwzględniają one jednak nieliniowych zależności między zmienną objaśnianą a zmiennymi objaśniającymi. W celu wychwycenia zależności nieliniowych stosuje się m.in. sztuczne sieci neuronowe. W przypadku szeregów czasowych powszechnie stosowane jest autoregresyjne podejście, które jako zmienne wejściowe przyjmuje również przeszłe wartości zmiennej prognozowanej. Szersze wyjaśnienie podstawowych metod prognozowania szeregów czasowych zostało przedstawione w dalszej części rozdziału.

Regresja wieloraka

Regresja wieloraka to technika uczenia się polegająca na dopasowaniu funkcji liniowej do wielu zmiennych niezależnych. Zakładając, że mamy n rekordów danych oraz p zmiennych objaśniających, model można sformułować w następujący sposób:

$$y = X\beta + \varepsilon \tag{2.3}$$

gdzie:

- y - wector$(n \times 1)$ danych objaśnianych,
- X - macierz$(n \times (p+1))$ zmiennych objaśniających,
- β - wector $((p+1) \times 1)$ współczynników,
- ε - wector $(n \times 1)$ błędów.

•

Parametry równania regresji nie są znane i muszą być wyznaczone na podstawie zbioru danych. Estymaty współczynników regresji β wyznacza się metodą najmniejszych kwadratów, minimalizując resztkową sumę kwadratów między zaobserwowanymi odpowiedziami w zbiorze danych, a odpowiedziami przewidywanymi przez przybliżenie liniowe. Resztkową sumę kwadratów można sformułować jako:

$$min||X\beta - y||^2 \tag{2.4}$$

Wieloraka regresja liniowa przyjmuje następujące założenia:

- istnieje liniowa zależność między zmiennymi niezależnymi a zmienną prognozowaną;
- liczba obserwacji musi być większa bądź równa liczbie parametrów (współczynniki oraz wyraz wolny);
- wariancja błędów jest taka sama dla wszystkich obserwacji;
- błędy mają rozkład zbliżony do rozkładu normalnego.

Regresja grzbietowa

Regresja grzbietowa (ang. ridge regression) [32] rozwiązuje niektóre problemy regresji wielorakiej, w szczególności zjawisko współliniowości występujące w modelach zawierających dużą liczbę zmiennych niezależnych.Regresja grzbietowa to szczególny przypadek

regularyzacji – tichonowa, gdzie na wszystkie parametry nałożona jest jednakowa kara [28]. Współczynniki regresji grzbietowej minimalizują resztkową sumę kwadratów, na którą nałożona została kara:

$$min||X\beta - y||^2 + \alpha||\beta||^2 \tag{2.5}$$

gdzie:

- $\alpha > 0$ jest parametrem złożoności, który kontroluje stopień kary.

Regresja grzbietowa zmniejsza wartości współczynników modeli w kierunku wartości zerowych lub w kierunku wzajemnych wartości. Jednakże w przypadku, gdy zmienne niezależne nie posiadają wspólnej skali, zmniejszanie wartości parametrów jest niewspółmierne. Dwie oddzielne zmienne posiadające różne skale będą przejawiać różny wpływ na nakładaną karę, ponieważ kara jest wyznaczana jako suma kwadratów wszystkich współczynników. By uniknąć wspomnianego problemu zmienne niezależne są poddawane standaryzacji lub normalizacji tak, aby wariancja zmiennej wynosiła 1 [14].

Autoregresja z niezależnymi zmiennymi wejściowymi

Autoregresja jest modelem, który wykorzystuje wartości zmiennej prognozowanej z poprzednich kroków czasowych jako danych wejściowych do modelu. Dodatkowe cechy wprowadzane do modelu, niebędące opóźnieniami zmiennej prognozowanej, nazywane są zmiennymi niezależnymi. utoregresja z niezależnymi zmiennymi wejściowymi jest sformułowana w następujący sposób:

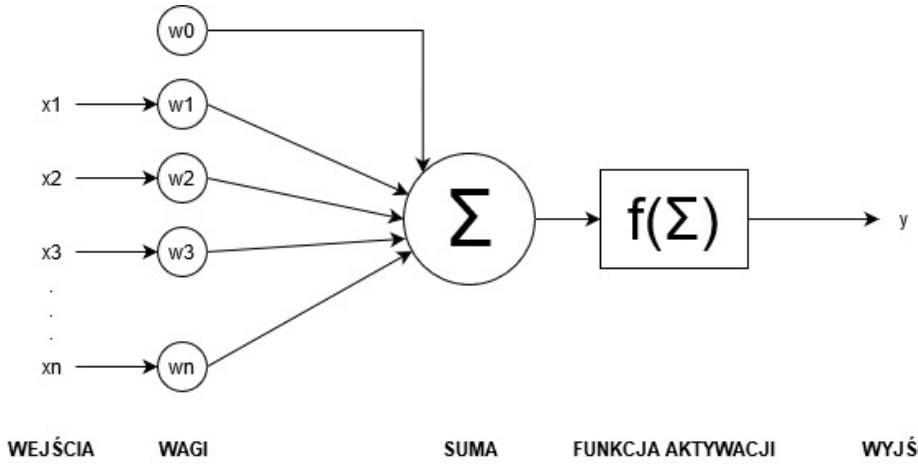
$$y_t = \gamma + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j x_j + \varepsilon_t \tag{2.6}$$

gdzie *gamma* to wyraz wolny, y_t - wartość obserwowana w czasie t,, α_i - współczynniki autoregresji, β_j - współczynniki zmiennych niezależnych. By znaleźć współczynniki α, β oraz γ wykorzystywana jest metoda najmniejszych kwadratów lub regresja ridge.

Sztuczne Sieci Neuronowe SSN

Sztuczna sieć neuronowa to algorytm uczenia maszynowego bazujący na modelu ludzkiego neuronu. Ludzki mózg składa się z milionów neuronów, wysyła i przetwarza informacje w postaci sygnałów elektrycznych i chemicznych. Neurony są połączone specjalną strukturą zwaną synapsami. Synapsy pozwalają neuronom na przekazywanie sygnałów, natomiast z dużej liczby symulowanych neuronów powstają sieci neuronowe.

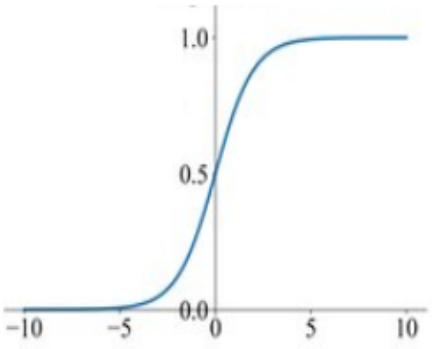
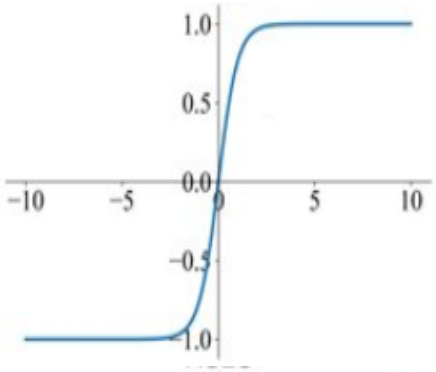
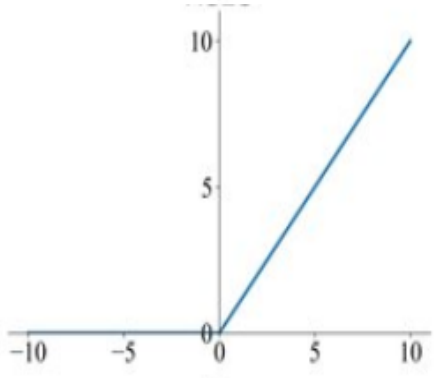
Podstawową jednostką SSN jest neuron, do którego wejść doprowadzane są sygnały dochodzące z neuronów warstwy poprzedniej. Każdy sygnał mnożony jest przez odpowiadającą mu wartość liczbową zwaną wagą. Wpływa ona na percepcję danego sygnału wejściowego i jego udział w tworzeniu sygnału wyjściowego przez neuron. Waga może być pobudzająca – dodatnia lub opóźniająca – ujemna; jeżeli nie ma połączenia między neuronami to waga jest równa zeru. Zsumowane iloczyny sygnałów i wag stanowią argument funkcji aktywacji neuronu. Wartość funkcji aktywacji jest sygnałem wyjściowym neuronu i propagowana jest do neuronów warstwy następnej. Funkcja aktywacji przybiera jedną z trzech postaci: skokowej, liniowej lub nieliniowej.



Rysunek 2.3: Model pojedynczego neuronu.

Wybór funkcji aktywacji zależy od rodzaju problemu, jaki stawiamy przed siecią do rozwiązania. Dla sieci wielowarstwowych najczęściej stosowane są funkcje nieliniowe, gdyż neurony o takich charakterystykach wykazują największe zdolności do nauki zależności występujących w danych wejściowych.

Najczęściej stosowana jest funkcja sigmoidalna zwana też krzywą logistyczną (przyjmuje ona wartości pomiędzy 0 a 1) oraz funkcja Rectified Linear Unit (ReLu) przedstawiona po raz

funkcja aktywacji	wzór	wykres
Sigmoid	$h_\theta(x) = \frac{1}{1+e^{-}\theta^T x}$	
Hyperbolic Tangent	$h_\theta(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	
Rectified Linear Unit	$h_\theta(x) = max(0,x)$	

pierwszy przez [30].]. Graficzna reprezentacja modelu pojedynczego neuronu przedstawiona jest na wykresie 2.3, a wartość wyjściowa neuronu wyznaczana jest wedle równania 2.7.

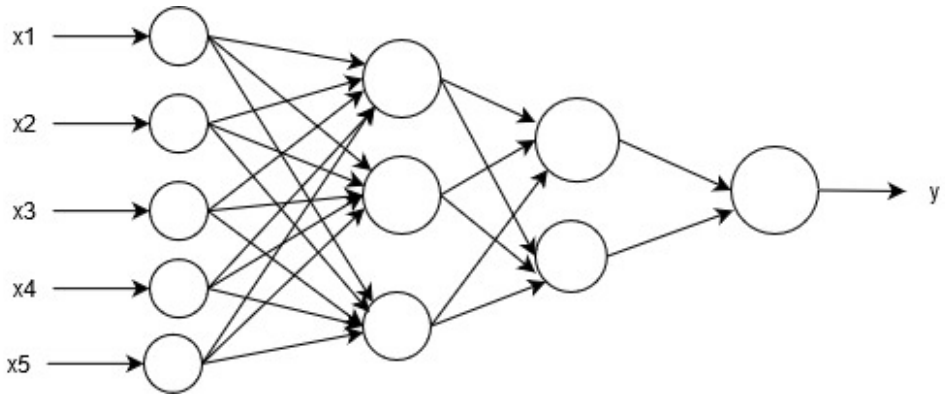
$$y = f(\sum_{i=1}^{n} w_i x_i + w_0) \tag{2.7}$$

Sieć wielowarstwowa powstaje przez odpowiednie połączenie wielu neuronów. Neurony tworzą warstwy, które połączone tworzą sieć. Sztuczna sieć neuronowa składa się z następujących warstw:

- warstwa wejściowa - warstwa z neuronami reprezentującymi dane wejściowe modelu,
- warstwa ukryta,

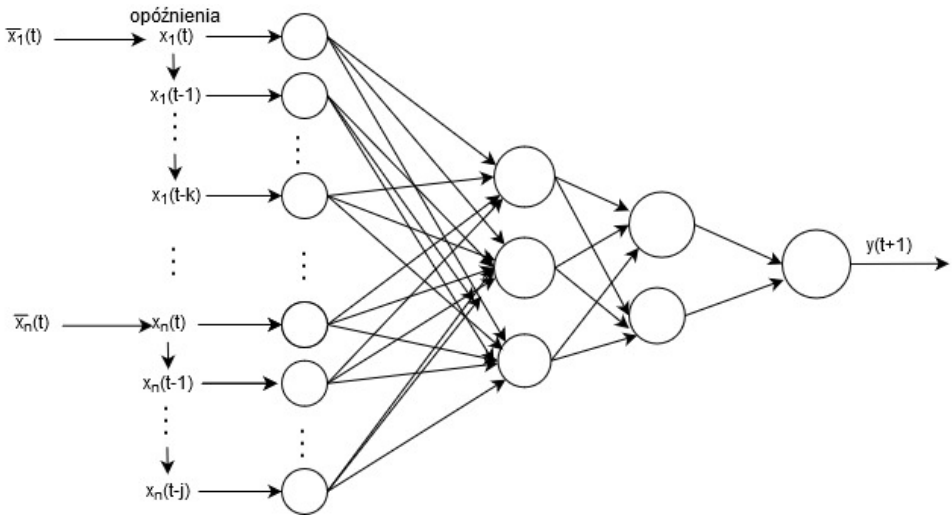
- warstwa wyjściowa - warstwa reprezentująca zmienne prognozowane.

Schemat przykładowej SSN z pięcioma cechami wejściowymi, dwoma warstwami ukrytymi oraz jedną zmienną prognozowaną przestawiony jest na wykresie 2.4.



Rysunek 2.4: Model sztucznej sieci neuronowej.

Proces trenowania sieci neuronowych jest procesem [27], podczas którego konieczny jest dobór szeregu parametrów definiujących sieć, m.in. funkcja aktywacji, metoda inicjalizacji wag sieci, stosowanie regularyzacji orazalgorytm optymalizujący wartości wag sieci. Do istotnych czynników wpływających na właściwości sieci neuronowej i jakość przybliżania zależności zawartych w prezentowanych danych treningowych należy także jej struktura, rozumiana zarówno jako liczba warstw ukrytych, jak i liczba neuronów w poszczególnych warstwach.



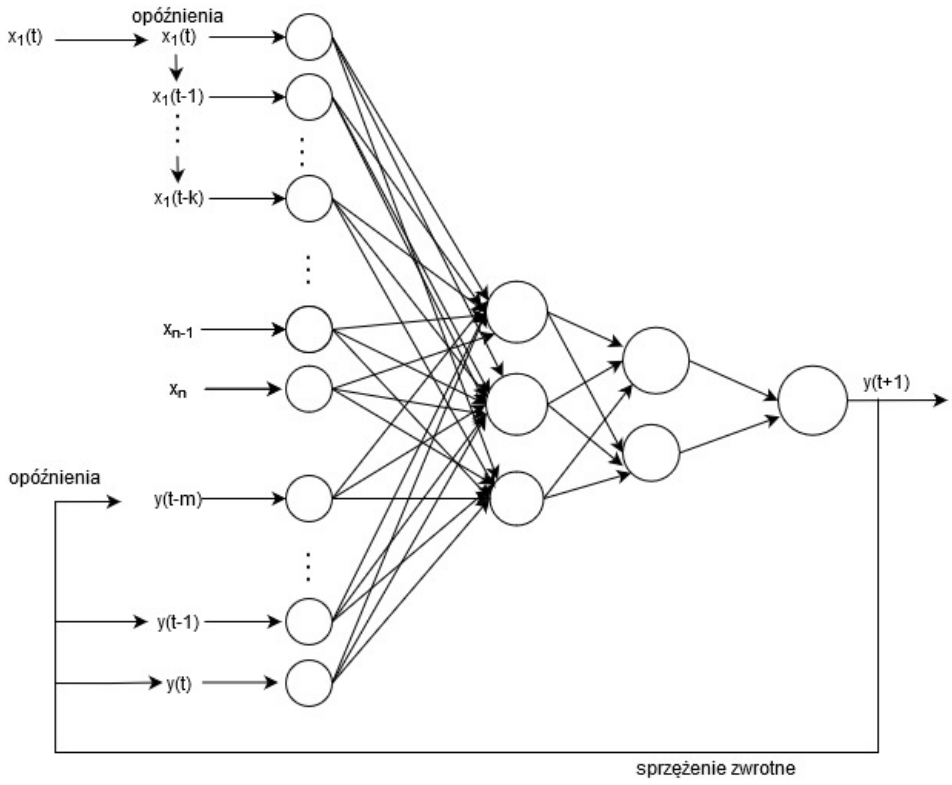
Rysunek 2.5: Model sztucznej sieci neuronowej z opóźnianymi wejściami w czasie.

Dane wejściowe sztucznej sieci neuronowej w przypadku prognozowania szeregów czasowych, często zawierając dane opóźnione w czasie, tzn. wyznaczając wartości

prognozowane dla kroku $t+1$ wykorzystywane są wartości zmiennych zależnych zarówno dla kroku t, jaki i kroków poprzednich $t-1, t-2, ..., t-m$, gdzie maksymalna wartość opóźniania zmiennych dobierana jest w sposób analityczny. Na rysunku 2.5 przedstawiona jest sieć ze zmiennymi zewnętrznymi $x1, x2, ..., xn$ które przed wprowadzeniem do pierwszej warstwy sieci są opóźniane. Należy mieć na uwadze, że wybrane opóźnienia dla każdej zmiennej zewnętrznej mogą być odmienne, różne mogą być także maksymalne opóźnienia zmiennych, wartości te dobierane są na etapie selekcji cech modelu według zdefiniowanego kryterium.

Sztuczne Sieci Neuronowe z autoregresyjnym wejściem

Analogicznie do autoregresyjnego modelu ze zmiennymi niezależnymi tworzone są modele SSN z wejściem autoregresyjnym.



Rysunek 2.6: Model sztucznej sieci neuronowej z autoregresyjnym wejściem

W takim przypadku warstwa wejściowa powiększana jest o wartości zmiennej prognozowanej z poprzednich kroków czasowych. Przykład takiej sieci przedstawiony jest na rysunku 2.6, a prezentowana sieć posiada jeden neuron wyjściowy, w związku z czym może on służyć do sekwencyjnego wyznaczania kolejnych wartości prognozy pojedynczej wartości prognozowanej, np. w przypadku prognozy zapotrzebowania na ciepło wyznaczana

jest wartosc prognozy na następną godzinę $t+1$ a następnie przekazywana jako wartość wejściowa do wyznaczania kolejnego kroku prognozy, tj. wartości dla godziny t+2. Sprzężenie zwrotne zaznaczone na rysunku 2.6 umożliwia przesyłanie i zapamiętywanie prognozowanych wartości tak, aby dane wejściowe mogły wykorzystywać różne wartości opóźnień zmiennej prognozowanej. Sieć przedstawiona na rysunku 2.6 posiada także wejścia zewnętrzne z wartościami opóźnianymi ($X1$) oraz wartościami aktualnymi.

2.3 Wskaźniki jakości prognoz

Szczególnie istotnym etapem prognozowania jest weryfikacjia prognoz i ocena ich jakości decydująca o ewentualnym wykorzystaniu modelu prognostycznego w praktyce. Oceny jakości prognoz dokonuje się poprzez analizę błędów prognozy. Zgodność wartości prognozowanych z danymi empirycznymi określa sie za pomocą szeregu wskaźników, z których najważniejsze zostały przedstawione w nieniejszym rozdziale. Dokładność predykcji wyznaczana jest na podstawie różnych zestawów danych, w szczególności ważna jest ocena wskaźników jakości prognoz dla danych treningowych (danych wykorzystanych do dopasowania współczynników modelu) oraz danych testowych (danych stosowanych wyłączonie do oceny jakości prognoz). Porównanie wskaźników jakości prognoz dla danych treningowych i testowych pozwala na ocenę modelu pod kątem przetrenowania (ang. overfitting) i daje podstawę do założenia, że model będzie uzyskiwał analogiczne wyniki na nowych danych.

Podstawowym wskaźnikiem stosowanym podczas analizy modeli prognostycznych jest **współczynnik determinacji** R^2 który wyznaczany jest wedle formuły 2.8.

$$R^2 = 1 - \frac{\sum_{t=0}^{n-1}(\hat{y}_t - \bar{y}))^2}{(y_t - \bar{y}))^2} \tag{2.8}$$

gdzie:

- y_t -wartość zmiennej Y w momencie lub okresie t,
- $\bar{y}$ - średnia wartość zmiennej Y w momencie lub okresie t,
- $\hat{y}_t$ - prognozowana wartość zmiennej Y w szeregu czasowym o długości n.

Współczynnik determinacji R^2 jest miarą dopasowania liniowego modelu regresji do danych rzeczywistych i informuje o tym, jaka część zmienności zmiennej objaśnianej w próbie pokrywa się z korelacjami ze zmiennymi zawartymi w modelu. Gdy parametry modelu są szacowane metodą najmniejszych kwadratów, współczynnik R^2 przybiera wartości z przedziału [0,1], przy czym im wyższa jest jego wartość, tym lepsze dopasowanie modelu.

Średni błąd absolutny (ang. mean absolute error) w czasie t wyznaczany jest wedle formuły:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{2.9}$$

[oznaczenia jak we wzorze (2.8)]

Wartość błędu informuje o ile średnio, w okresie predykcji, rzeczywiste wartości zmiennej prognozowanej odchylają się, co do bezwzględnej wartości, od prognoz. Im niższą wartość przyjmuje MAE tym lepsze jest dopasowanie modelu.

Średni absolutny błąd procentowy w czasie t (ang. mean absolute percentage error) wyznaczany jest wedle formuły:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{2.10}$$

[oznaczenia jak we wzorze (2.8)]

Wartość błędu informuje o średniej wielkości błędów prognoz wyrażonych w procentach. Wartości MAPE pozwalają porównać dokładność prognoz otrzymywanych przy zastosowaniu różnych modeli.

Stosowanie wskaźników dokładności predykcji wyznaczających wartości średnie pozwala na pominięcie problemu wartości odstających, jednocześnie pozwalając na zwiększanie błędów w przypadku ich wystąpienia [78].

2.4 Przedziały ufności

Przedziały ufności są sposobem na wyrażenie niepewności związanej z prognozą, reprezentują zakres, w którym z pewnym prawdopodobieństwem będzie leżała wartość, którą prognozujemy. Im węższe są przedziały ufności, tym dokładniejsza jest prognoza.

W przypadku prognozowania szeregów czasowych przedział ufności jest mocno zależny od horyzontu predykcji. Innymi słowy, im bliższy jest horyzont prognoz, tym pewniejsze są prognozy. Korzyścią z wyznaczania przedziałów ufności jest możliwość oceny jej niepewności. Analizując wykres, a w szczególności szerokość przedziałów, można sprawdzić jakość modelu. Szerokie przedziały ufności świadczą o dużej niepewności generowanych przez model prognoz. Taka graficzna reprezentacja jest bardziej intuicyjna w szczególności w przypadku regularnego korzystania z generowanych przez model prognoz.

Analiza przedziałów ufności bazuje na wyznaczonych błędach predykcji zgodnie z równaniem 2.11.

$$e = y_t - \hat{y}_t \tag{2.11}$$

[oznaczenia jak we wzorze (2.8)]

Zakres przedziałów ufności zależny jest od stopnia skomplikowania modelu oraz liczby zmiennych wejściowych. Wielkość przedziałów ufności dla modelu bez wejść autoregresyjnych, w przypadku gdy błędy prognoz definiowane są wedle równania 2.11 oraz prezentują rozkład normalny, jest taka sama dla każdego kroku czasowego w zakresie horyzontu predykcji. Innymi słowy, zakres błędu predykcji w przypadku modelu bez autoregresji jest stały niezależnie od godziny horyzontu predykcji. Rozpiętość przedziałów ufności dla takiego przypadku wyznaczana jest wedle zależności 2.12.

$$< y^* - t_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}} : y^* - t_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}} > \tag{2.12}$$

gdzie:

- $t_{\frac{a}{2}}$ - dystrybuanta rozkładu t-studenta dla $1-a$ i n-1 stopni swobody
- σ - błąd standardowy błędów predykcji e
- $\sqrt{n}$ - wielkość próbki danych

W przypadku wyznaczania przedziałów ufności dla modelu z wejściem autoregresyjnym konieczne jest wyznaczenie rozpiętości przedziału ufności dla każdej kolejnej godziny

predykcji. Wynika to z faktu, iż w takim przypadku wartości prognozowane stanowią wejście w kolejnych krokach czasowych horyzontu predykcji. Kolejne godziny predykcji kumulują, więc błąd prognozy pojawiający się w poprzednich godzinach horyzontu predykcji. Wartości rozpiętości przedziału ufności dla każdej godziny wyznaczane są indywidualnie wedle równania 2.12.

Rozdział 3

Prognozowanie zapotrzebowania na cieplo w sieciach ciepłowniczych

W tym rozdziale przedstawione zostaną dwie metody wyznaczania zapotrzebowania na ciepło w sieciach ciepłowniczych. Następnie przedstawione zostaną obecnie stosowane podejścia oraz metody prognozowania zapotrzebowania na ciepło w kontekście zarówno całych sieci ciepłowniczych, jak i indywidualnych odbiorców.

3.1 Metody wyznaczania zapotrzebowania na ciepło

Przedstawiając temat zapotrzebowania na ciepło kluczowym askepktem rzutującym na sposób ujęcia tematów oraz osiągane wyniki jest definicja pojęcia zapotrzebowania na ciepło.

Pierwsze podejście zakłada, że zapotrzebowanie na ciepło w sieci to ilość ciepła dostarczanego do sieci przez źródło ciepła. Wartość tą w przypadku sieci, której medium to ciepła woda, wyznaczana jest według następującej formuły:

$$ZC(t) = c_p T_1(t) m_1(t) - c_p T_2(t) m_2(t) \tag{3.1}$$

gdzie:

- c_p - ciepło właściwe wody,
- T_1 - temperatura wody wprowadzanej do sieci,
- T_2 - temperatura wody powrotnej z sieci,

- m_1 - masa wody wprowadzanej do sieci,
- m_1 - masa wody powrotnej z sieci.

Wyznaczając zapotrzebowanie na ciepło w ten sposób uwzględniamy w prognozie bezpośrednią konsumpcję ciepła przez odbiorców oraz straty ciepła w sieci (dystrybucja ciepła). Zaletą tej metody jest jej prostota wyznaczania i duża pewność, wymagany jest pomiar maksymalnie 4 parametrów, aby możliwe było obliczenie wyniku. Wadą metody jest uwzględnianie w wyniku strat ciepła w sieci, które nie są zależne od zachowań konsumentów, będących przedmiotem zainteresowania podczas tworzenia prognozy, a jedynie od warunków pogodowych oraz stanu sieci. W przypadku modernizacji sieci, np. wymiany części rur skutkującej znaczącym spadkiem strat ciepła stworzone modele prognostyczne tracą dokładność prognozy. W takim przypadku konieczne jest zebranie nowych danych funkcjonowania sieci i rekalibracja modelu. Ponowny proces zbierania danych wymaga odczekania reprezentatywnego horyzontu czasu, co w przedstawianym w niniejszej pracy modelu wynosi minimalnie pełny rok pracy sieci.

Drugą możliwą metodą wyznaczania zapotrzebowania na ciepło jest sumowanie konsumpcji ciepła przez odbiorców końcowych na podstawie liczników zużycia ciepła w węzłach cieplnych. W takim przypadku zapotrzebowanie na ciepło wyznaczane jest według następującej formuły:

$$ZC(t) = \sum_{i=1}^{n} Q_i(t) \tag{3.2}$$

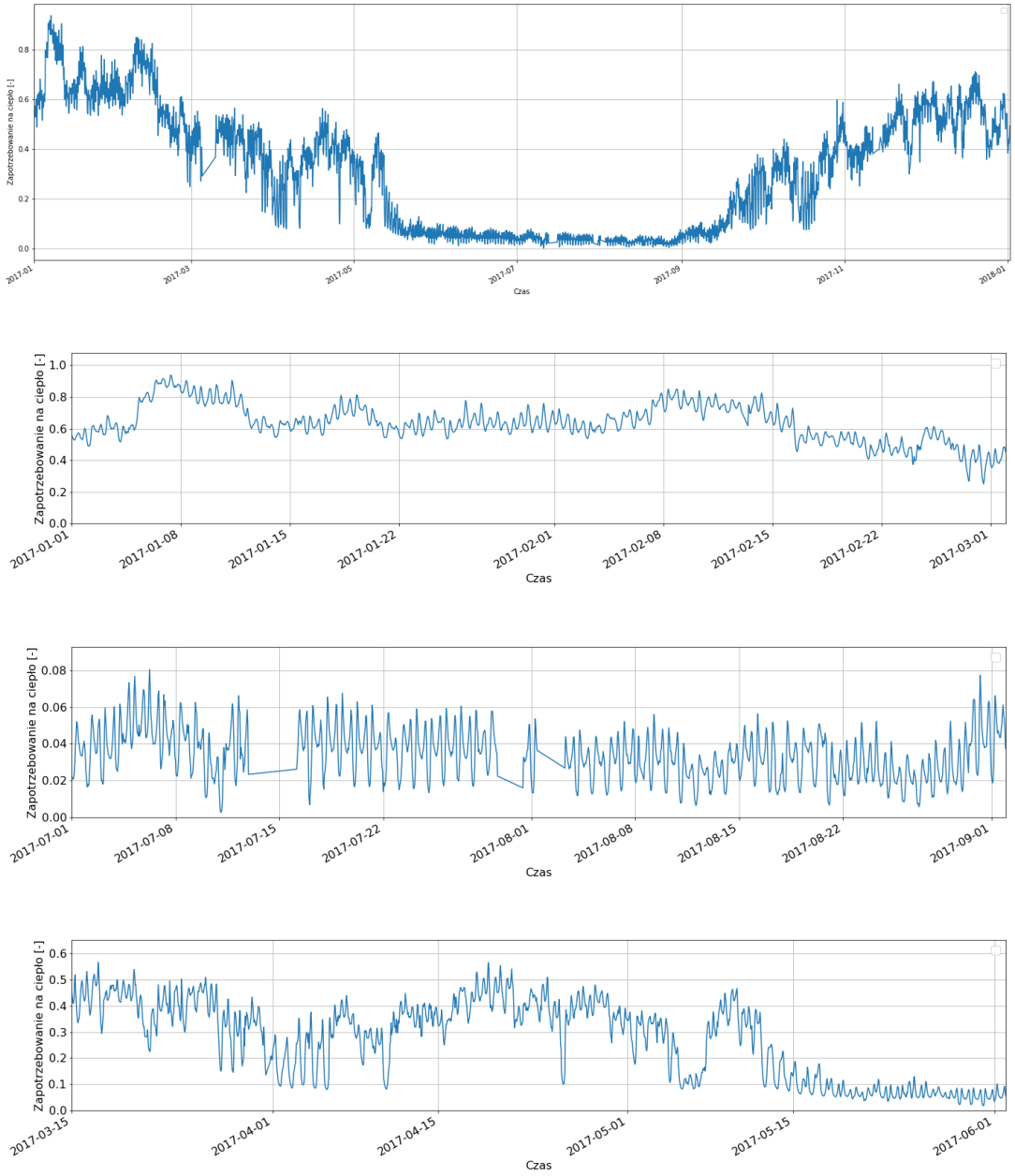
gdzie:

- n - liczba odbiorców w sieci,
- Q_i - konsumpcja ciepła przez odbiorcę i.

Taka metoda wyznaczania zapotrzebowania na ciepło pozwala na pominięcie strat ciepła w sieci powstałych w wyniku dystrybucji ciepła. Trudność w precyzyjnym stogowaniu tej metody, w szczególności w przypadku dużych sieci ciepłowniczych dostarczających ciepło do tysięcy odbiorców, jest konieczność zbierania danych pomiarowych z każdego węzła cieplnego. Oznacza to konieczność instalacji inteligentnych liczników, które odczytują, zapisują i przesyłają dane ze wskazaną rozdzielczością, ponadto metoda ta wymaga stworzenia metody pozwalającej na pominięcie problemu chwilowego braku danych z wybranej grupy/wybranych odbiorców pojawiających się w wyniku np. awarii licznika.

3.2 Sezonowość zapotrzebowania na ciepło

Sieci ciepłownicze zapewniają dostarczenie ciepłej wody użytkowej i ciepła do odbiorców końcowych w celach grzewczych. Z uwagi na występowanie w trakcie roku okresów niewymagających ogrzewania budynków charakterystyka przebiegu oraz poziom zapotrzebowania na ciepło jest zależny od pory roku.



Rysunek 3.1: Zapotrzebowanie na ciepło w zależności od sezonu.

Przeskalowane do przedziału [0,1] zapotrzebowanie na ciepło dla warszawskiej sieci ciepłowniczej dla całego roku kalendarzowego oraz wybranych pór roku przedstawione

jest na wykresie: 3.1., zauważalne jest, że dla okresu letniego zapotrzebowanie na ciepło charakteryzuje się powtarzalnym przebiegiem z występującymi nieznacznymi fluktuacjami, a w okresie grzewczym przebieg jest bardziej dynamiczny i nieregularny. Największe, nieregularne zmiany w wysokości zapotrzebowania występują w okresie przejścia z sezonu letniego do pełnego sezonu grzewczego oraz w okresie przejścia z pełnego sezonu grzewczego do sezonu zimowego. W tych okresach następuje stopniowe uruchamianie lub wyłączanie wymienników centralnego ogrzewania, co jest bezpośrednią przyczyną bardzo dużych dobowych amplitud zapotrzebowania na ciepło.

Z uwagi na przedstawione różnice w charakterystyce zapotrzebowania na ciepło w zależności od pory roku, wskazane jest, aby analiza możliwości prognozowania zapotrzebowania na ciepło również rozgraniczała wyniki między wymienione okresy, tj. sezon letni, sezon zimowy (grzewczy) oraz sezon przejściowy.

3.3 Przegląd literatury dotyczącej prognozowania zapotrzebowania na ciepło

Analizując temat prognozowania zapotrzebowania na ciepło, nalezy zwrócić uwagę na następujące aspekty:

- sposób wyznaczania wartości zapotrzebowania na ciepło, aspekt ten opisany został w rozdziale 3.1;
- wielkość sieci dla której wyznaczana jest prognoza, aspekt ten opisany został w rozdziale 3.1;
- okres dla którego tworzone są modele czy prognozowane jest zapotrzebowanie na ciepło w całym roku, czy wybranych sezonach, aspekt ten opsiany został w rozdziale 3.2;
- horyzont oraz krok czasowy predykcji modeli.

Ponadto ważnym czynnikiem w kontekście budowania systemów bazujących na metodach uczenia maszynowego jest klasa i typ stosowanych modeli oraz wykorzystane dane wejściowe. W zależności od klasy modelu czas trenowania oraz wymagane parametry sprzętowe mogą znacząco od siebie odbiegać. Skomplikowane modele, np. głębokie sztuczne sieci neuronowe

wymagają znacznego czasu trenowania sieci oraz doboru hiperparametrów, natomiast zaletą tych modeli jest większa dokładność precyzji prezentowana w wielu badaniach w szczególności w przypadku gdy relacja między zmienną zależną a zmiennymi niezależnymi jest silnie nieliniowa. Mniej czasochłonne obliczeniowo modele, takie jak wieloraka regresja liniowa przy zmniejszonej dokładności precyzji pozwalają na skrócenie czasu trenowania modeli. Ważne jest w kontekście implementacji modeli w rzeczywistym systemie dobranie modelu prognostycznego pozwalającego na uzyskanie jak najwyższych wartości dokładności precyzji przy dostępnych ograniczeniach sprzętowych oraz dostępnym czasie trenowania modeli czy wyznaczania prognoz.

W dalszej części rozdziału przedstawione zostały prezentowane w literaturze naukowe metody prognozowania zapotrzebowania na ciepło ze szczególnym wyróżnieniem aspektów wymienionych na wstępie tego rozdziału.

3.3.1 Metoda wyznaczania zapotrzebowania na ciepło

Przedstawione w literaturze modele prognostyczne zapotrzebowania na ciepło w głównej mierze odnoszą się do wartości wyznaczania zapotrzebowania na ciepło w sieci, stosując formułę 3.1i są ukierunkowane na zwiększenie efektywności pracy źródeł ciepła. Przykładowo autor w [26] przedstawia model wyznaczania dziennej wariacji produkcji ciepła bazującą na analizie szeregów czasowych, którego wyniki wykorzystywane są do optymalizacji pracy magazynów ciepła. Metoda 3.1 wyznaczania zapotrzebowania na ciepło jest stosowana w szczególności w przypadku, gdy analizie poddana jest duża sieć ciepłownicza składająca się z setek lub tysięcy odbiorców ciepła, np. miejska sieć ciepłownicza w Rydze [8, 58], Lublanie[59] czy Aarhus [21].

Modele prognostyczne zapotrzebowania na ciepło wyznaczanego przy pomocy formuły 3.2 pokrywają obszar badań, gdzie analizowani są wybrani odbiorcy ciepła, przy czym liczba badanych odbiorców często liczona jest maksymalnie w dziesiątkach. Autorzy w [55] przedstawili analizę optymalizacji pracy sieci ciepłowniczej składającej się z 25 węzłów cieplnych bazującą na prognozie zapotrzebowania na ciepło dla odbiorców końcowych. Autor badań w [35] przedstawia analizę sieci w Szwecji składającej się z 5 tys. odbiorców, jednakże opisane są indywidualne modele dla wybranych odbiorców, natomiast analizę możliwych metod prognozowania dla pojedynczych budynków przedstawili autorzy w [65, 75].

Prezentowane w literaturze naukowej badania nie obejmują prognozy zapotrzebowania na ciepło rozumianej jako suma konsumpcji ciepła przez odbiorców, tj. wyznaczanego zgodnie z formułą 3.2. W szczególności zauważalny jest brak badań dotyczący dużych sieci ciepłowniczych, gdzie liczba odbiorców liczona jest w tysiącach, natomiast wraz ze znacznym postępem w cyfryzacji sieci ciepłowniczych oraz instalacji telemetrii ukierunkowanej na powstawanie inteligentnych sieci ciepłowniczych takie badania stają się możliwe do przeprowadzenia. Przedstawiony w niniejszej pracy system prognostyczny uzupełnia obszar badań prognozowania zapotrzebowania na ciepło o analizę możliwości prognozowania zapotrzebowania na ciepło bazującego na pomiarach z węzłów cieplnych w opozycji do prognoz generowanych na podstawie pomiarów ze źródeł ciepła. Zmiana ta stanowi jeden z elementów pozwalających na transformację obecnych systemów ciepłowniczych w kierunku systemów czwartej generacji.

3.3.2 Typy i klasy modeli prognostycznych

Prezentowane w literaturze modele prognostyczne zapotrzebowania na ciepło w większości wykorzystują modele regresji liniowej oraz sztucznych sieci neuronowych. Sieć ciepłownicza w Rydze została przeanalizowana w [58] i [8]. W badaniu [58] zastosowano średnią z dwóch metod: nieliniowej autoregresyjnej sieci neuronowej z dodatkowym wejściem oraz regresji wielorakiej, uzyskując dokładność predykcji MAPE równą 4,76 proc. dla 24-godzinnych prognoz na zbiorze testowym, który zawierał dane z grudnia. W kolejnym badaniu [8]autorzy przetestowali nieliniową autoregresyjną sieć neuronową, uzyskując dokładność produkcji dla 14. zimowych dni na poziomie 6,02 proc. dla parametru MAPE. Różne liniowe i nieliniowe modele prognostyczne dla sezonu zimowego oparte na sieciach neuronowych są porównywane w [59] gdzie wysunięto wniosek, że modele sieci neuronowej są nieco lepsze niż modele liniowe. Bardziej zaawansowany model uczenia maszynowego dla okresów przejściowych został przedstawiony w [38] prezentującym rekurencyjną sieć neuronową. Natomiast w [37] randomizowane drzewo decyzyjne jest porównywane z sztuczną siecią neuronową ze sprzężeniem zwrotnym na korzyść sieci neuronowej, osiągając dokładność predykcji MAPE na poziomie 11,7 proc. dla okresu styczeń–marzec. Model sezonowej autoregresyjnej zintegrowanej średniej ruchomej (SARIMA) został przetestowany na zbiorze danych testowych z ostatnich 20 pełnych tygodni roku w [25] dla prognozy zapotrzebowania na ciepło w mieście Espoo w Finlandii. Modele prognostyczne oceniono na podstawie godzinowych

danych pogodowych (temperatura zewnętrzna i prędkość wiatru) oraz godzinowych danych zużycia ciepła. Wyniki obliczone jako średnia ze statystyk dla 20-tygodniowego horyzontu prognostycznego, gdzie prognoza dokonywana była na jeden tydzień, osiągnęła wartość 0,9154 dla współczynnika determinacji R kwadrat i 6,86 proc. dla skorygowanego średniego bezwzględnego błędu procentowego (AMAPE). Autorzy sugerują możliwość oceny modeli dla poszczególnych budynków pod warunkiem dostępności automatycznych liczników odczytujących pomiary klientów.

Prowadzone są również badania dotyczące prognozy zapotrzebowania na ciepło dla poszczególnych odbiorców końcowych z wykorzystaniem metod modelowania analogicznych do opisanych powyżej. Sieć neuronowa przedstawiona w [75] została opracowana w celu prognozowania zapotrzebowania na ciepło na Politechnice Warszawskiej, natomiast sieć neuronowa autoregresyjna opisana w [65] z dodatkowym wejściem została opracowana w celu stworzenia modelu dla trzech budynków komercyjnych, uzyskując średnią dokładność 4-godzinnych predykcji dla danych zawierających pełny rok kalendarzowy na poziomie 3,2 proc. parametru MAPE. Modele uczenia maszynowego, takie jak maszyna wektorów nośnych (SVR), drzewa regresyjne, jednokierunkowa sztuczna sieć neuronowa (FFNN) i wieloraka regresja liniowa (MLR) są analizowane w [35] w celu stworzenia prognozy dla indywidualnych odbiorców w Szwecji. Uzyskane wyniki w zależności od analizowanej podstacji dla 24-godzinnego horyzontu predykcji wynoszą od 5,56 proc. do 21,54 proc. Autorzy w [21] przeanalizowali SVR, FFNN i MLR do prognozowania sieci ciepłowniczych w Aarhus. W [70] autorzy przedstawiają metody prognozy dla systemów Karlshamm i Rottne różnymi metodami, przy czym metoda sztucznej sieci neuronowej przewyższa inne analizowane, uzyskując dokładność predykcji MAPE 8,08 proc. dla okresu od sierpnia do lutego.

3.3.3 Zakres wyników badań

Badania prezentowane w literaturze w większości skupiają się na prognozowaniu zapotrzebowania na ciepło w sezonie grzewczym [58, 59, 37, 70, 21] z uwagi na znacznie wyższy poziom zapotrzebowania, a w związku z tym dużo większe straty w sieci, niektóre wyniki prezentowane są dla całego roku, jednak nie pojawia się tam analiza dokładności predykcji w rozbiciu na sezony, jedynie autor [22] przedstawia wykresy dokładności modelu

w podziale na miesiące, natomiast podział na lato i sezon grzewczy zastosował autor [25]. Prezentowane w literaturze modele prognostyczne dla indywidualnych odbiorców również skupiają się w głównej mierze na sezonie grzewczym [35]. Pojedyńcze prace analizują szerszy horyzont, autor [65] przedstawia wyniki dla całego roku jednak bez podziału na sezony, uwypuklając jednocześnie niską dokładność modeli dla sezonu letniego spowodowaną niedopasowaniem urządzeń pomiarowych. Natomiast autor [55] przedstawia modele dla wybranych tygodni reprezentujących wszystkie okresy roku (wiosna, lato, jesień zima) jednakże brakuje prezentacji wyników dokładności predykcji. Prezentowane modele stanowią dane wejściowe do optymalizatora i tylko w takim kontekście są prezentowane.

3.3.4 Wnioski wynikające z przeglądu literatury

Posumowując, głowne wnioski można streścić w następujących punktach:

1. Możliwe jest precyzyjne prognozowanie zapotrzebowania na ciepło z wykorzystaniem modeli uczenia maszynowego.

2. Najlepszym modelem prognostycznym często są sztuczne sieci neuronowe zarówno dla indywidualnych odbiorców ciepła, jak i całej sieci ciepłowniczej.

3. Nie ma dostępnych badań dla dużych sieci ciepłowniczych, gdzie zapotrzebowanie na ciepło wyznaczane jest z wykorzystaniem pomiarów indywidualnie dla każdego węzła cieplnego.

4. Badnia dotyczące indywidualnych odbiorców analizują jedynie pojedyńcze węzły cieplne, a badania analizujące większą liczbę odbiorców w sieci są bardzo ograniczone. Brak takich badań nie pozwala na generalizację problemów czy efektywności modeli przedstawianych w literaturze.

5. Prezentowane wyniki dla modeli prognostycznych odnoszą się do pełnego sezonu grzewczego, czyli okresu, gdy uruchomione są wszystkie wymienniki centralnego ogrzewania, podczas gy okres przejściowy jest okresem newralgicznym. Ponadto zakres danych testowych nie obejmuje całego sezonu a wybrane odcinki, np. tygodnie lub dni.

6. Modele prezentowane w literaturze wykorzystują maksymalnie cztery cechy związane z pogodą. Zawsze wykorzystywana jest temperatura zewnętrzna jako podstawowa cecha wejściowa.

7. Prezentowane modele prognostyczne wyznaczające zapotrzebowanie na ciepło dla całej sieci uzyskują dokładność predykcji dla 24-godzinnego horyzontu dla parametru MAPE w przedziale 4,76 proc.–11,7 proc.. Przytoczone wyniki predykcji odnoszą się do zbiorów testowych bazujących na danych z okresu sezonu grzewczego, brakuje natomiast analizy dokładności predykcji dla danych z okresu sezonu letniego oraz przejściowego.

Przedstawiony w tym rozdziale przegląd badań dotyczących prognozowania zapotrzebowania na ciepło potwierdza zasadnosć badań przedstawionych w niniejszej rozprawie. Prezentowany system w szczególności uzupełnia obszar badań nad możliwościami prognozowania zapotrzebowania na ciepło o:

1. Metodologię oraz analizę możliwści wyznaczania zapotrzebowania na ciepło na podstawie danych telemetrycznych z węzłów cieplnych wraz z walidacją oraz kompensacją brakujacych danych.

2. Selekcję najlepszego modelu wyznaczającego prognozy zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej.

3. Analizę dokładności predyckji dla krótkiego (24 godziny) oraz średniego (120 godzin) horyzontu predykcji dla całego roku kalendarzowego, tj. sezonu letniego, sezonu grzewczego oraz sezonu przejściowego.

4. Analizę dokładności predyckji modelu po wdrożeniu w rzeczywistym systemie.

Rozdział 4

Metodologia budowy systemów opartych na technikach uczenia maszynowego

Uczenie maszynowe (UM) to technika analizy danych, której celem jest umożliwienie komputerom wykonywania zadań, analogicznych jakie wykonują ludzie, przez proces uczenia się. Algorytmy uczenia maszynowego wykorzystują metody obliczeniowe do „uczenia się" informacji bezpośrednio z danych, pomijając odgórnie określone relacje, np. w postaci równań fizykalnych. W rzeczywistości uczenie maszynowe przeprowadza analizę danych, wykorzystując zestaw zaimplementowanych instrukcji stosowanych przez różnorodne algorytmy zaprojektowane do podejmowania decyzji i/lub prognozowania [49, 33]. Uczenie maszynowe jest złożone, sposób działania zależy w dużej mierze od wykorzystywanych algorytmów oraz dostępnych danych. Podstawowe koncepcje UM to:

- Uczenie nadzorowane (ang. supervised learning). Celem uczenia nadzorowanego jest znalezienie modelu odwzorowującego dane wejściowe na dane wyjściowe na podstawie dostarczonych przez człowieka przykładowych par wejścia-wyjścia. Takie algorytmy zasilane są dużą ilością danych uczących. W uczeniu nadzorowanym każdy przykład to para składająca się z obiektu wejściowego i żądanej wartości wyjściowej. Uczenie nadzorowane wykorzystuje się głównie do modelowania procesów technicznych.

- Uczenie bez nadzoru (ang. unsupervised learning). Jest to koncepcja polegająca na znalezieniu funkcji opisującej strukturę danych nieetykietowanych, czyli takich które nie zostały wcześniej połączone w pary wejście-wyjście. Główną cechą odróżniającą tę metodę od metody nadzorowanej jest to, że nie ma prostego sposobu oceny jakości

algorytmu, ponieważ przykłady podawane jako dane wejściowe są nieoznakowane. Metoda wykorzystywana jest głównie do statystyki oraz do estymacji funkcji gęstości prawdopodobieństwa.

Tabela 4.1 przedstawia szeroko stosowane techniki UM oraz przykłady ich praktycznych zastosowań. Systemy UM mogą rozwiązywać problemy, którymi nie można zarządzać ręcznie ze względu na ogromną ilość danych, które muszą zostać przetworzone.

Tabela 4.1: Techniki uczenia maszynowego.

Algorytm UM	**Opis**	**Przykłady zastosowania**
Klasyfikacja	określenie, na podstawie danych wejściowych, oddzielnych klas, do których należy rekord	filtrowanie wiadomości na spam i nie-spam, wykrywanie oszustw finansowych, personalizacja treści, wykrywanie wad produkcyjnych, oznaczenia zgłoszeń do pomocy technicznej,
Regresja	prognoza na podstawie danych wejściowych wartości dla każdego rekordu	prognozowanie zachowań rynku, prognozowanie zapotrzebowania, zarządzanie ryzykiem, zarządzanie zasobami, prognozowanie pogody, prognozowanie w sporcie
Detekcja anomali	identyfikacja obiektów, zdarzeń lub obserwacji, które nie są zgodne z oczekiwanym wzorcem lub innymi elementami w zbiorze danych	detekcja anomali w urządzeniach, sieciach i systemach
Grupowanie	dzielenie danych na grupy (klastry) na podstawie prawdopodobieństwa	segmentacja klientów, identyfikacja wzorców zachowań, grupowanie obserwowanych epicentrów trzęsień ziemi w celu identyfikacji niebezpiecznych obszarów
Rekomendacje	prognozowanie alternatywnych wyborów użytkowników	rekomendacje produktów, rekrutacje, randkowanie online, rekomendacja treści
Imputacja	uzupełnianie wartości brakujących danych	brakujące dane od klientów, dane spisowe, niekompletne raporty medyczne

Ponadto techniki UM mogą wyznaczyć subtelne zależności i relacje pojawiające się w danych, które są niemal niemożliwe do wykrycia podczas manualnej analizy. Natomiast połączenie wielu subtelnych zależności może stanowić silną podstawę do generowania trafnych

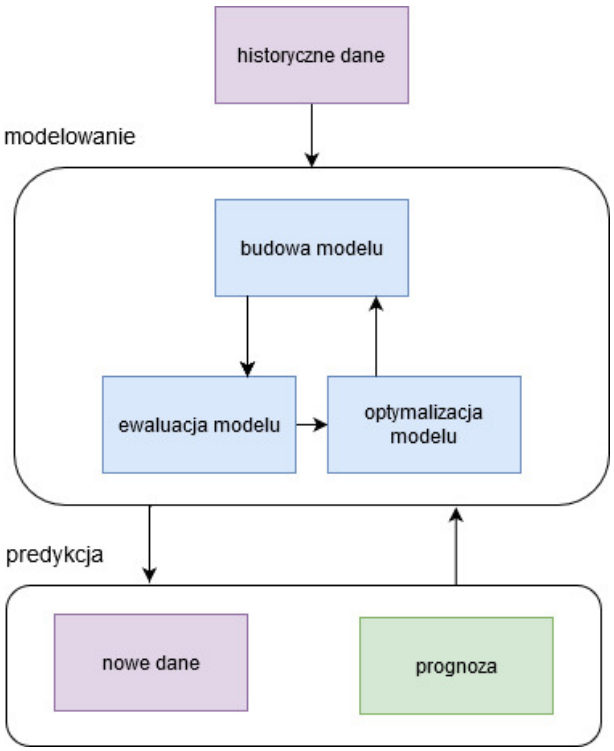
prognoz. Dlatego że proces uczenia się z danych, a następnie wykorzystywania zdobytej wiedzy do informowania o zdarzeniach z przyszłości jest cenny, UM bardzo szybko stało się silnikiem napędowym współczesnej ekonomii bazującej na danych (ang. data-driven). Przewaga systemów bazujących na technikach UM w porównaniu z manualną analizą oraz zakodowanymi na stałe regułami czy prostą analizą statystyczną w głównej mierze odnosi się do takich obszarów jak:

- *dokładność* - UM wykorzystuje dane do eksploracji silnika pozwalającego na podejmowanie optynalnych decyzji w zależności od badanego probelmu. Wraz z zwiększaniem ilości dostępnych danych dokładność algorytmu może sie zwiększać;
- *automatyzacja* - wraz ze stworzeniem rozwiązania problemu, modele UM mogą uczyć się nowych zależności automatycznie. Pozwala to na osadzenie algorytmów UM bezpośrednio w zautomatyzowanym procesie.
- *szybkość* - stworzone modele UM wraz z pojawieniem się nowych danych wyznaczają odpowiedzi w ciągu kilku sekund, co pozwala na rekcję w czasie rzeczywistym;
- *konfigurowalność* - modele UM są budowane na podstawie indywidualnych danych reprezentujących badany problem, natomiast rozwiązanie może zostać skonfigurowane tak, aby optymalizować obszar, który najbardziej wpływa na dane przedsiębiorstwo;
- *skalowalność* - wraz ze wzrostem przedsiębiorstwa, UM jest w stanie przetworzyć zwiększoną ilość zbieranych danych. Niektóre algorytmy UM są w stanie analizować ogromne ilości zebranych danych jednocześnie na wielu urządzeniach w chmurze.

Wysokie wymagania jakościowe algorytmów zarówno w zakresie projektowym, jak i programistycznym pozwalają na implementację zróżnicowanych funkcjonalności algorytmów UM. W ciągu ostatniej dekady UM oraz głębokie UM wniosło wiele w ogromną liczbę dziedzin badawczych i inżynieryjnych, takich jak eksploracja danych (ang. data mining) [79], medycyna [40, 52], nauki o ziemi [45], energetyka [77, 44], sieci inteligentne [61], komunikacja [11], rozpoznawanie obrazu [39], śledzenie obiektów [18] i innych. Popularność oraz efektywność takich systemów możliwa jest dzięki znacznemu postępowi, jaki dokonał się na przestrzeni ostatnich lat w obszarze ilości zbieranych danych oraz mocy obliczeniowej komputerów.

System bazujący na technikach UM to zbiór komponentów zdolnych do nauki powiązań występujących w danych oraz wykorzystania zdobytej wiedzy do wyznaczenia predykcji

określającej przyszłość [67]. Budowa rzeczywistych systemów bazujących na technikach UM wymaga realizacji następujących etapów przedstawionych na wykresie 4.1: przygotowanie danych, budowa modelu, ewaluacja modelu, optymalizacja modelu, oraz predykcja na podstawie nowych danych. Kolejność realizacji poszczególnych etapów jest naturalna, jednakże większość aplikacji do uczenia maszynowego w świecie rzeczywistym wymaga wielokrotnego powtarzania każdego kroku w procesie iteracyjnym. W dalszej części rozdziału przedstawione zostały kluczowe aspekty tychże etapów w odniesieniu do systemów opierających się na modelach prognostycznych.



Rysunek 4.1: Proces tworzenia systemu opartego na technikach uczenia maszynowego[13]. Na podstawie przygotowanych danych historycznych budowany jest model UM. W kolejnym etapie zdolności predykcyjne modelu podlegają ewaluacji, a następnie optymalizacji tak, aby dopasować działanie modelu do oczekiwanych wyników. Gotowy model pozwala na wyznaczanie predykcji dla nowych danych.

4.1 Definicja problemu

Przed przystąpieniem do budowy systemu konieczne jest zdefiniowanie podstawowych założeń, które rzutują na stosowane algorytmy uczenia maszynowego. Proces definicji założeń wspierają następujące pytania:

- Co chcemy osiągnąć, budując ten algorytm?

- Jak możemy zdefiniować stawiany problem w ujęciu uczenia maszynowego? Jaki typ uczenia maszynowego będzie tutaj odpowiedni: uczenie z nauczycielem, uczenie bez nauczyciela czy uczenie przez wzmacnianie?
- Co ma być wynikiem algorytmu? Jaki typ uczenia maszynowego rozwiązuje postawiony problem? Regresja, klasyfikacja czy rekomendacja?
- Jakie dane są wymagane do rozwiązania probelmu? Jaki minimalny zakres danych jest konieczny do zebrania?
- W jaki sposób oceniony będzie sukces modelu? Jakie metryki dokładności modelu będa stosowane? Jaka dokładność modelu jest akceptowalna?

Proces budowania systemu ML jest procesem ciągłym i w trakcie jego tworzenia, na podstawie bieżących wyników, odpowiedzi na przedstawione wyżej pytania mogą zostać zmodytfikowane. Przykładowo, może się okazać że zebrane dane nie pozwalają na stworzenie odpowiednio dobrego modelu. W takiej sytuacji konieczne jest ponowne zdefiniowanie i zebranie danych, zmiana definicji problemu lub stwierdzenie o niemożności rozwiązania przedstawianego problemu z wykorzystaniem narzędzi ML.

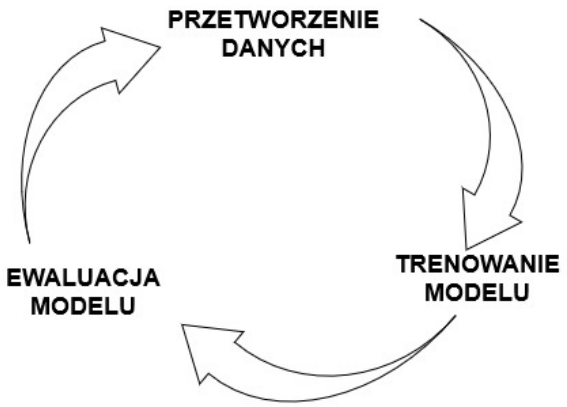
Wynikiem opisywanego w tym rozdziale kroku budowania systemów ML jest definicja problemu w ujęciu uczenia maszynowego, zbiór danych oraz definicja metryk i ich oczekiwanej dokładności.

4.2 Modelowanie

Proces modelowania składa się z trzech obszarów, przedstawionych na wykresie 4.2, które wykonywane są iteracyjne do momentu uzyskania akceptowalnych rezultatów.

4.2.1 Przetwarzanie danych

Jakość danych oraz ilość użytecznych informacji w nich zawartych są kluczowymi czynnikami decydującymi o jakości algorytmów uczenia maszynowego. Dane niepoddane żadnej transformacji nazywane są danymi surowymi i wymagają odpowiedniego przetworzenia, aby możliwe było stworzenie i wykorzystanie stosownych algorytmów. W poszczególnych etapach przetwarzania danych wykonuje się następujące procedury [60]:



Rysunek 4.2: Faza modelowania podczas tworzenia systemu ML.

Zebranie oraz przygotowanie odpowiedniej ilości i typów danych jest procesem wymagającym wnikliwego i indywidualnego podejścia do każdego procesu zarówno podczas budowy modułu, jak i implementacji gotowych rozwiązań. Konieczne jest wstępne określenie wymaganych typów danych, np. dane pogodowe, dane z mierników oraz ich zakresu czasowego. Proces ten jest procesem kluczowym w świetle jakości oraz funkcjonalności systemu i wymaga głębokiego przeanalizowania badanego zagadnienia. Często, w szczególności w przypadku procesów inżynieryjnych, wymagana jest także wiedza ekspercka z danej dziedziny. Wynikiem końcowym tego etapu jest zunifikowana baza danych o zdefiniowanej strukturze i postaci.

W dominującej większości przypadków dane przetwarzane są do postaci tabelarycznej w taki sposób, aby kolejne wiersze reprezentowały kolejne rekordy pomiarowe (w przypadku szeregów czasowych, kolejne wiersze reprezentują kolejne pomiary w czasie), natomiast kolumny wartość cechy dla danego rekordu/kroku w czasie. Przykład zestawu danych przedstawiony został na wykresie 4.3, w takim zestawie danych pojedyncze kolumny reprezentują taki sam typ danych, natomiast poszczególne wiersze zazwyczaj zawierają różne typy danych. W tabeli przedstawionej na wykresie 4.3 zawarte są dane ciągłe (*temperatura*), kategoryczne (*zachmurzenie, status CO*) oraz tekstowe (*nazwa węzła*). Taki zestaw danych musi zostać przetworzony do postaci, która jest akceptowana przez algorytmy UM.

Walidacja danych surowych wykonywana jest w celu wykrycia oraz usunięcia ze zbioru danych surowych błędnych rekordów. Pojęcie błędnych rekordów danych definiowane jest indywidualnie dla każdego zbioru danych, klasycznie są to pomiary o bardzo dużej lub niskiej wartości. Ponadto w przypadku procesów fizykalnych możliwe jest stworzenie eksperckich

Cechy w kolumnach

Pomiar	temperatura	zachmurzenie	status CO	nazwa węzła
2015-01-01 01:00:00+00:00	-1,5	3	Włączony	węzeł_A
2015-01-01 02:00:00+00:00	3,2	4	Wyłączony	węzeł_B

rekordy w wierszach

Rysunek 4.3: Przykład tabularycznego zestawu danych wejściowych.

reguł pozwalających na walidację danych, np. bilans masowy.

Zarządzanie brakującymi danymi jest istotne w konteście tworzenia niezawodnych systemów oraz precyzyjnych algorytmów, ponieważ braki w danych w sposób istotny wpływają na efektywność trenowania modeli UM. Najczęściej stosowane metody zarządzania brakujacymi danymi to [60]:

- Usówanie ze zbioru danych wierszy zawierających brakujące dane. Rozwiązanie to jest szczególnie nieefektywne w przypadku występowania dużego odsetka brakujących danych, odrzucenie tych danych spowoduje znaczny spadek możliwości predykcyjnych modelu. Ponadto w przypadku szeregów czasowych usuwanie rekordów zawierających brakujące dane, powoduje dzielenie zbioru danych na wiele podzbiorów o ciągłej linii czasu, co uniemożliwia efektywne trenowanie modeli autoregresyjnych.

- Uzupełnienia brakujących danych pewną wartością, np. zerem, wartością sąsiednią (z poprzedniego kroku czasowego), średnią wartością lub medianą danej kolumny/cechy. W takim podejściu uzupełniane dane będą znacznie odbiegać od rzeczywistych wartości, jednakże pozwalają na zachowanie ciągłości danych, w przypadku gdy bardziej zaawansowane metody nie są możliwe do wykorzystania, np. w przypadku bardzo dużych zbiorów danych.

- Uzupełnienia brakujących danych średnią wartością z najbliższych pomiarów/rekordów. Rozwiązanie najczęściej stosowane ze względu na prostotę i większą dokładność w porównaniu z metodami przedstawionymi wyżej. W zależności od rozkładu wartości danej cechy stosowana jest również mediana, gdyż wartość średnia jest bardziej podatna na wartości odstające.

- Uzupełnienia brakujących danych wartością prognozowaną. W takim wypadku budowane są proste algorytmy UM takie jak regresja liniowa.

Inżynieria cech to praktyka polegająca na wykorzystaniu matematycznych przekształceń surowych danych wejściowych do tworzenia nowych cech będących potencjalnym wejściem do modeli UM. Innymi słowy, inżynieria cech jest procesem transformującym zebrane surowe dane do postaci, która może być bezpośrednio wykorzystana przez algorytmy UM. Często zależności wykryte przez algorytmy UM pozwalają na wyznaczanie dokładnej prognozy, jednakże interpretowalność modeli może być niska, mowa tu o na przykład o tzw. modelach typu czarna skrzynka (ang. black box), gdzie podawane są tylko dane wejściowe i zmienna zależna natomiast sam sposób wyznaczania prognozy jest niezrozumiały/ukryty wewnątrz modelu, przykładem takich modeli są głębokie sieci neuronowe. W takich przypadkach bardziej wartościowe może być zaprojektowanie nowych cech, które będą reprezentowały proces generalizujący analizowany zestaw danych oraz intuicyjne powiązanie między danymi surowymi a zmienną zależną. Podczas tworzenia nowych cech wykorzystywana jest w głównej mierze wiedza ekspercka w zakresie analizowanego systemu i zebranych danych, która pozwala na przekształcenie obecnej wiedzy w zakresie analizowanego systemu w zbiór danych wejściowych do modelu UM.

Dodatkowym aspektem inżynierii cech jest transformacja danych kategorycznych, czyli danych, z których każda może przyjąć jedną ze skończonej liczby kategorii. Korzystając z algorytmów UM należy przekształcić wartości kategoryzujące na postać liczbową, gdyż algorytmy UM nie są dostosowane do analizowania innych typów danych, np. ciągów znaków. Powszechnie stosowane metody transformacji danych kategorycznych to [60]:

- mapowanie z ciągu znaków na postać całkowitoliczbową w przypadku danych kategorycznych porządkowych, czyli danych, których kategorie są możliwe do porównania, np. informacja o liczbie cylindrów w układzie kategoryzowana jako v3,v4,v8;

- kodowanie jeden do wielu (ang. one hot encoding) w przypadku danych kategorycznych nominalnych, czyli danych, których kategorie nie są możliwe do porównania, np. informacja o kraju pochodzenia. Kodowanie jeden do wielu zamienia kodowaną kolumnę na N-1 kolumn, gdzie N, to liczba kategorii nominalnych. Każda taka kolumna reprezentuje jedną kategorię i przechowuje wartości zero lub jeden informujące o przynależności do danej kategorii.

Kategoryczne dane binarne, tzn. przyjmujące wyłącznie dwie klasy, np. włączone/wyłączone nie wymagają kodowania.

Proces inżynierii cech jest szczególnie istotny w przypadku wykorzystywania danych nieustrukturyzowanych (tekst, obrazy, wideo, kroki czasowe), które w surowej postaci nie są możliwe do wykorzystania przez modele UM.

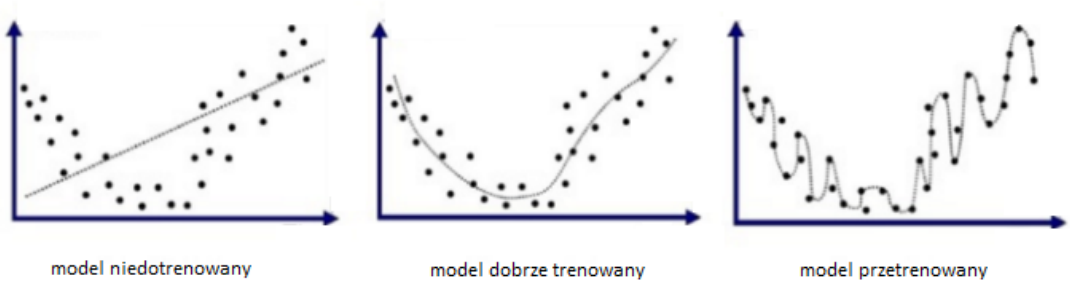
Podział zbioru danych na zbiór treningowy i testowy dokonywany jest po to, aby w przyszłości możliwa była ocena jakości modelu. W szczególności ważna jest ocena czy stworzony model dobrze generalizuje badane zjawisko, uzyskując podobne wyniki na danych niewykorzystanych w procesie trenowania modelu, tj. danych testowych. Analiza wyników prognoz w zbiorze treningowym i testowym pozwala na wykrycie dwóch zjawisk wpływających na osiąganie znacząco niższej precyzji modeli w chwili adaptacji do nowych danych:

- przetrenowania modelu (ang. overfitting),czyli zbyt sztywne dopasowanie modelu do konkretnych przykładów uczących (danych treningowych) bez możliwości generalizowania zależności na nowe dane. W takim przypadku model osiąga bardzo dobre wyniki tylko dla zbioru danych treningowych, natomiast dla nieznanych przypadków, np. zbioru danych testowych, model osiąga istotnie gorsze wyniki. Zjawisko przetrenowania występuje, gdy model jest zbyt skomplikowany w porównaniu do ilości lub zaszumienia danych uczących. Taki model uzyskuje bardzo dobre wyniki w obszarze zbioru treningowego, natomiast niektóre dane modelu nie zostają odwzorowane na zbiorze testowym;

- niedotrenowanie modelu (ang. underfitting) , które występuje gdy model jest zbyt prosty aby skutecznie nauczyć się zależności występujących w daych.

Graficzne przedstawienie zjawiska przetrenowania oraz niedotrenowania modelu przedstawione zostało na wykresie 4.4. Model niedotrenowany stanowi jedną linię, niedostatecznie reprezentując zależności występujące w danych, natomiast model przetrenowany jest wielomianem wysokiego stopnia, który zbyt mocno dopasował się do danych.

Metodologia podziału danych na zbiór treningowy i testowy nazywana jest walidacją krzyżową (ang. cros-validation), natomiast najbardziej powszechne metody walidacji krzyżowej to metoda podziału losowego (ang. hold-out) oraz metoda k-krotnej walidacji krzyżowej (k-fold cross validation).



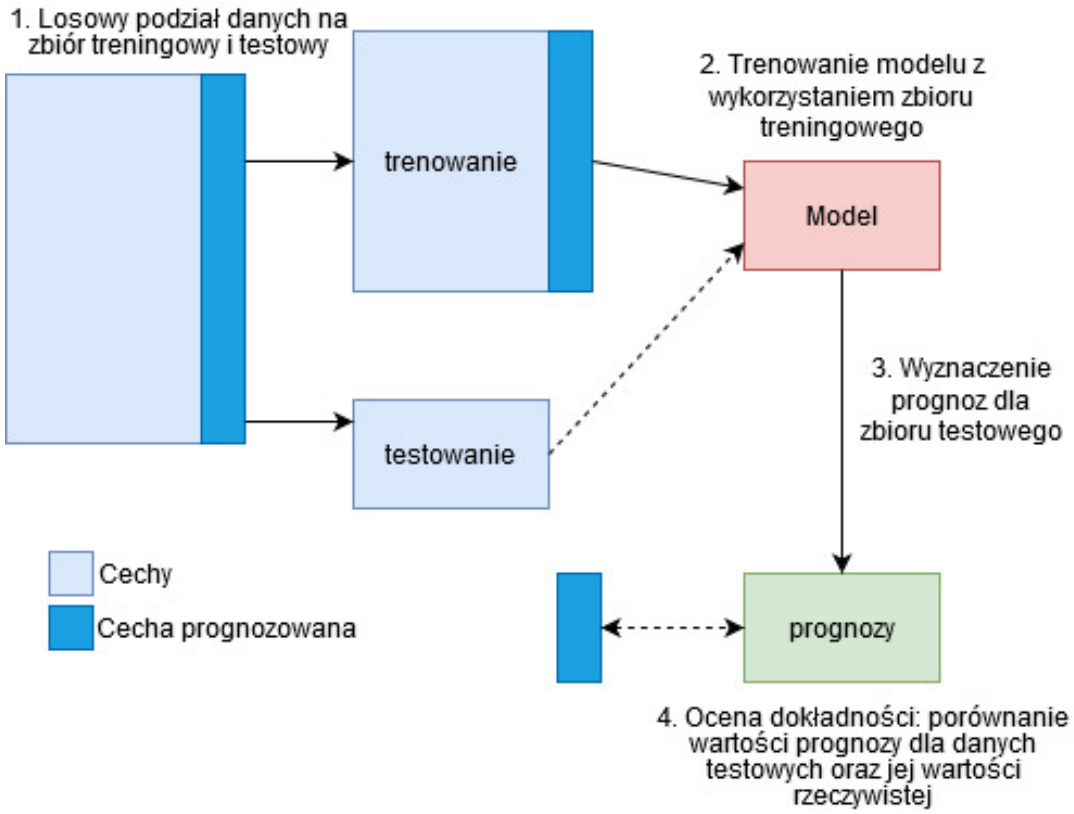
Rysunek 4.4: Przykłady niedotrenowania oraz przetrenowania modelu UM.

Metoda podziału losowego oddziela randomową część danych przypisanych jako dane testowe od procesu uczenia. Zbiór testowy stanowi od 20 do 40 proc. zestawu danych. Schemat metodologii trenowania oraz oceniania dokładności predykcji modelu przedstawiony został na wykresie 4.5. W pierwszym kroku następuje losowy podział danych na zbiór treningowy i testowy, następnie na postawie zbioru treningowego trenowany jest model UM, w kolejnym kroku z wykorzystaniem danych ze zbioru treningowego wyznaczane są prognozy, które finalnie porównane są z rzeczywistymi wartościami tejże cechy. Porównanie wartości prognozowanych oraz rzeczywistych zarówno dla zbioru treningowego, jak i testowego stanowi podstawę oceny modelu.

W przypadku modeli dla szeregów losowych, podział na zbiór treningowy i losowy nie jest dokonywany losowo z uwagi na możliwe zależności między następującymi po sobie pomiarami. Ponadto losowy podział w przypadku szeregów czasowych spowodowałby sytuację, w której dane z przyszłości stanowiłyby podstawę do prognozy zmiennej z przeszłości [72]. W takim wypadku stosuje się podział danych surowych na ciągłe odcinki czasu tak, aby dane treningowe czasowo następowały po danych testowych.

Metoda k-krotnej walidacji krzyżowej jest lepszym rozwiązaniem, jednakże bardziej wymagającym obliczeniowo. Metoda ta dzieli zbiór danych surowych na k rozłącznych podzbiorów. Dla każdego podzbioru model jest trenowany z wykorzystaniem wszystkich danych surowych z wyjątkiem danych z danego podzbioru, a następnie model jest wykorzystywany do wyznaczania predykcji dla danych z danego zbioru. Po przeanalizowaniu wszystkich k podzbiorów, prognozy dla każdego podzbioru są agregowane i porównywane z rzeczywistą wartością w celu oceny dokładności modelu. Metodologia k-krotnej walidacji krzyżowej została zobrazowana na wykresie 4.6.

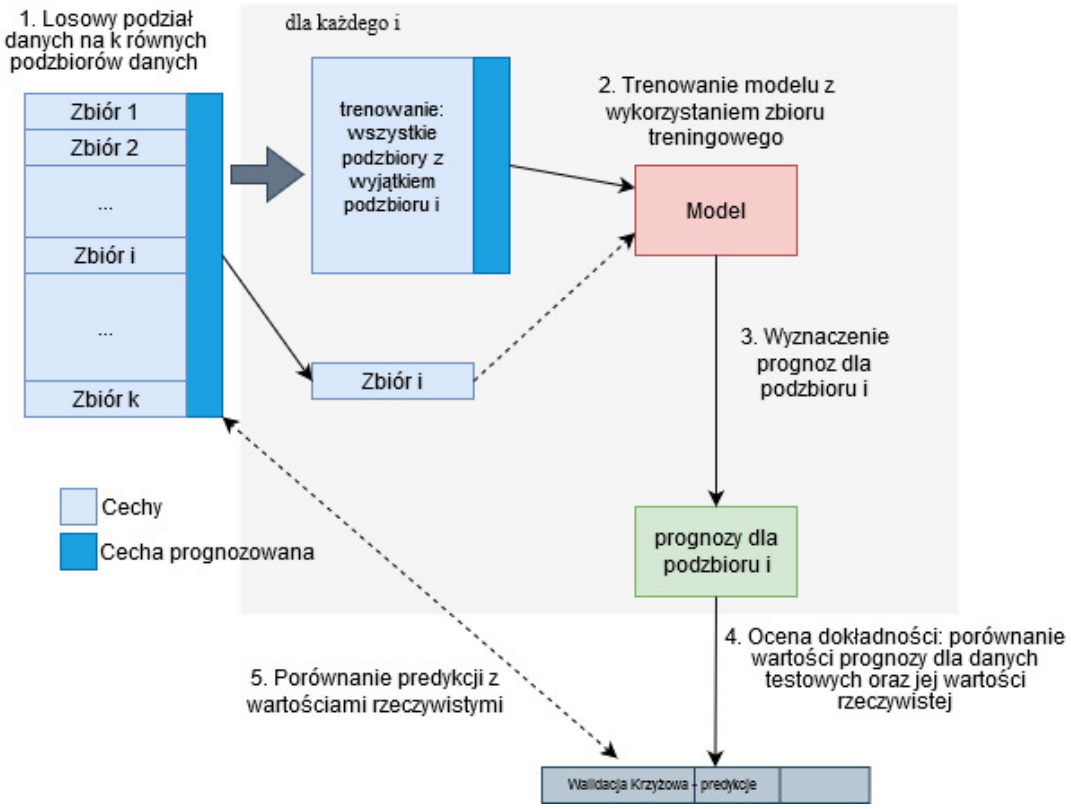
Walidacja krzyżowa pozwala na estymację dokładności modelu UM w przypadku



Rysunek 4.5: Trenowanie oraz ocena precyzji modelu z wykorzystaniem metody poziału losowego.

wykorzystania nowych danych (wdrożenie modelu), należy natomiast mieć na uwadze następujące aspekty:

- metoda walidacji krzyżowej zakłada, że dane treningowe stanowią reprezentatywną próbę danych z populacji. W przypadku wdrożenia modelu do predykcji na podstawie nowych danych należy upewnić się, że nowe dane pochodzą z tego samego źródła;
- stosując metodę walidacji krzyżowej należy upewnić się, że zbiór treningowy i testowy są zbiorami rozłącznymi to znaczy, że należy się upewnić, że w zbiorze treningowym nie znajdują się przesunięte lub poddane transformacji dane ze zbioru testowego. Zjawisko to nazywa się wyciekiem danych (ang. data lekage) i stanowi ważny aspekt trenowania modeli analizujących szeregi czasowe [5];
- wysoka liczba podziału na podzbiory zapewnia lepszą estymację błędu modelu, natomiast jednocześnie wydłużony zostaje czas trenowania. Literatura podaje 10-krotną walidację krzyżową jako optymalną [42].



Rysunek 4.6: Trenowanie oraz ocena precyzji modelu z wykorzystaniem metody k-krotnej walidacji krzyżowej.

4.2.2 Trenowanie modeli

Jak przedstawiono w rozdziale 2 wiele różnych algorytmów zostało stworzonych w kontekście prognozowania szeregów czasowych, dlatego ważnym punktem podczas budowy systemu jest wybór klasy algorytmu wykorzystanego do generowania prognoz. Podczas trenowania oraz ewaluacji modelu wykonuje się takie czynności jak: selekcja cech wejściowych do modelu, dobór hiperparametrów modelu.

Selekcji cech wejściowych do modelu. Główną zaletą modeli UM jest zdolność wykrywania zależności w przypadku bardzo dużej liczby cech i ilości danych wejściowych, ponieważ większa liczba cech może wspierać lepsze mapowanie cech wejściowych z cechą prognozowaną, strategiczne wydaje się dodawanie do zbioru danych jak największej liczby cech. Jednakże często liczba potencjalnych cech wejściowych jest nieproporcjonalnie większa w stosunku do liczby zebranych rekordów, czyniąc model podatnym na przetrenowanie. Liczba cech wejściowych wpływa także na czas i efektywność trenowania modelu. Selekcja cech jest wykonywana w celu [29]:

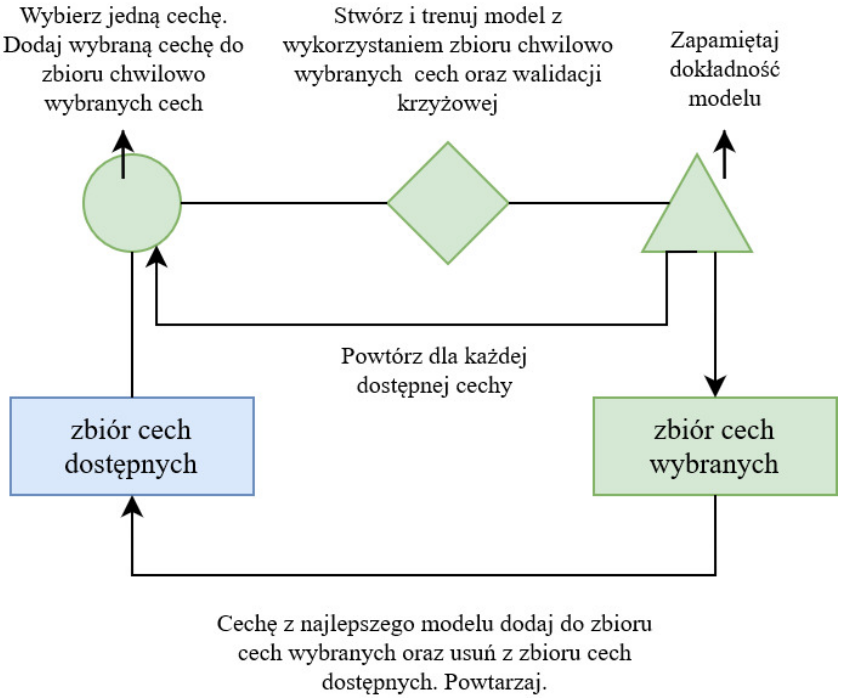
1. Uproszczenia modelu tak, aby ułatwić interpretację wyników oraz wpływu poszczególnych cech na wartości predykcji. W wielu przypadkach wykorzystania modeli UM ważniejsze jest zrozumienie, jak i dlaczego powstała dana prognoza natomiast sama wartość prognozy jest informacją drugorzędną.

2. Zwiększenia uogólnienia modelu przez ograniczenia przetrenowania modelu, a w konsekwencji pogorszenia jakości prognoz w przypadku wdrożenia modelu.

3. Skrócenia czasu trenowania modelu, który jest zazwyczaj zależny od liczby cech wejściowych, a tym samym rozmiaru macierzy danych.

Głównym założeniem podczas selekcji cech jest to, iż dane zawierają wiele cech, które są zbędne lub nieistotne, zatem mogą być usunięte bez ponoszenia dużych strat informacji.

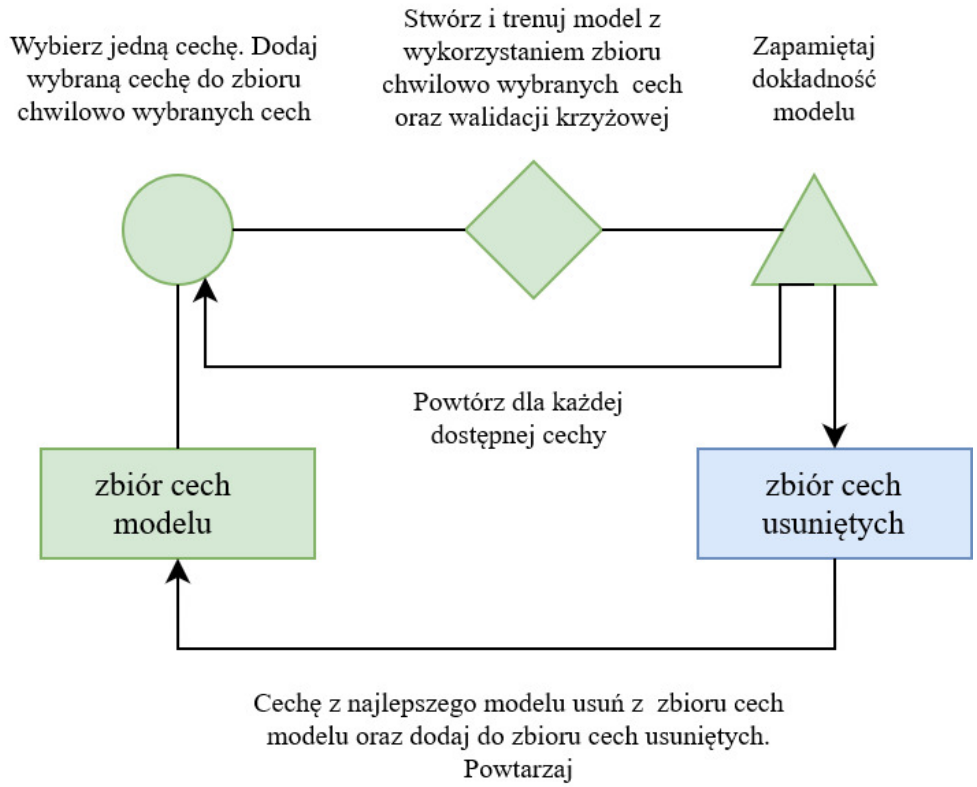
Jedną z podstawowych oraz najczęściej stosowanych selekcji cech jest **selekcja wstępująca,** w której istotne cechy dobierane są iteracyjne [29]. Idea algorytmu zakłada, że pierwotnym zbiorem cech wejściowych jest pusty algorytm, a w każdym kolejnym kroku iteracji dodaje się tę cechę, która daje najlepszą poprawę jakości dopasowania modelu. Algorytm selekcji wstępującej przedstawiony jest na wykresie 4.7. W zależności od liczby potencjalnych cech różna liczba modeli UM jest trenowana. Należy mieć na uwadze



Rysunek 4.7: Selekcja cech wejściowych do modelu z wykorzystaniem selekcji wstępującej i zstępującej.

to, że jedna z cech może być nieistotna w obecności innej istotnej cechy, z którą jest silnie skorelowana, dlatego selekcja wstępująca jest jedynie przybliżeniem najlepszego możliwego zestawu cech. Zaletą algorytmu jest możliwość śledzenia kolejności dodawanych cech, jednakże w przypadku licznego zbioru potencjalnych cech wejściowych czas doboru cech może znacząco się wydłużyć.

Wykres 4.8 przedstawia proces selekcji cech metodą selekcji zstępujacej, w tym przypadku pierwotny zbiór cech wejściowych stanowią wszystkie dostępne cechy, a wkolejnych iteracjach algorytmu cechy na podstawie, których model uzyskał najgorszą dokładność są eliminowane.



Rysunek 4.8: Selekcja cech wejściowych do modelu z wykorzystaniem selekcji zstępującej.

Dobór hiperparametrów modelu. Każdy model UM zawiera różny zestaw hiperparametrów, które kontrolują, jak model wykorzystuje dane treningowe do budowy modelu. Parametry te nie są wyuczane przez model uczenia maszynowego podczas procesu uczenia. Przykładowe hiperparametry dla algorytmów regresji to:

- regresja liniowa – brak;
- regresja grzbietowa – współczynnik kary α;

- sieci neuronowe – liczba ukrytych warstw i neuronów w każdej ukrytej warstwie, funkcja aktywacji.

W zależności od wybranego modelu po wstępnej weryfikacji zasadności stosowania modelu, tzn. uzyskaniu wyników dokładności modelu dla domyślnych wartości hiperparametrów, dokonuje się doboru optymalnych hiperparametrów, aby uzyskać lepszą jakość modelu. W zależności od złożoności modelu, ilości hiperparametrów oraz czasu trenowania modelu stosuje się różne techniki optymalizacji modelu, takie jak metoda siatki (ang. grid search) czy metoda losowania (ang. random search) [60].

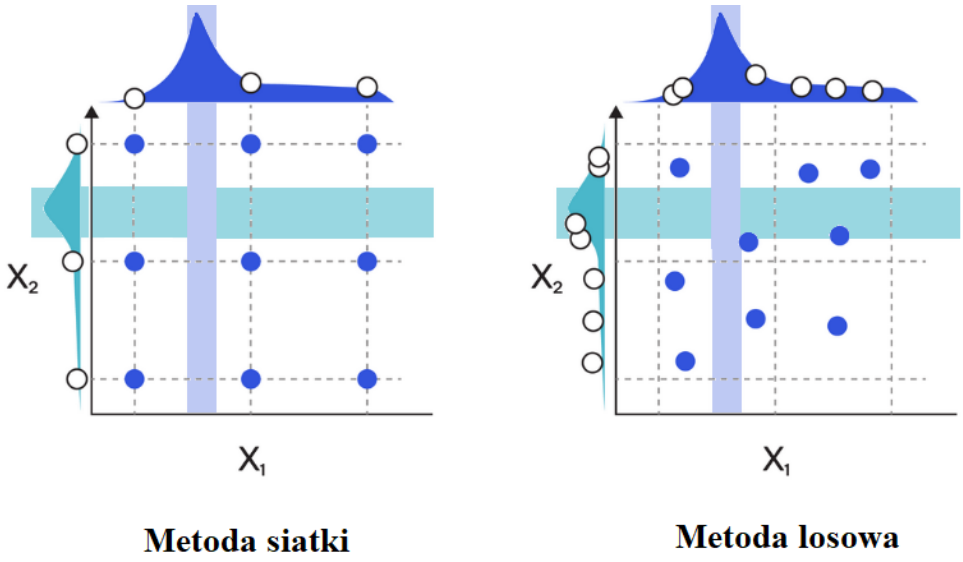
Metoda siatki polega na tworzeniu siatki parametrów, które będą testowane. W tej metodzie określany jest zakres wartości każdego parametru i jego skok, na tej podstawie budowana jest siatka parametrów, a następnie testowaniu każdego węzła siatki. Jeżeli przestrzeń parametrów jest mocno nieliniowa oraz jeden z parametrów ma niewielki wpływ na wyniki to poszukiwanie parametrów na siatce nie jest optymalnym rozwiązaniem. Regularność siatki powoduje, że możemy nie natrafiać na istotne maksima.

Metoda losowania polega na testowaniu losowych wartości parametrów, co pozwala zwiększyć szansę na znalezienie punktów charakteryzujących przestrzeń. Jest to metoda bardziej efektywna w porównaniu z metodą siatki [10]. Jednym z powodów, dla których wyszukiwanie metodą losową zazwyczaj jest leppsza od metody siatki, jest to, że często tylko kilka hiperparametrów ma znaczenie dla danego zbioru danych, a znalezienie optymalnych wartości dla dominujących hiperparametrów będzie miało większy wpływ niż uzyskanie optymalnej kombinacji wszystkich dostępnych hiperparametrów. Hiperparametry, które są ważne różnią się w zależności od danych treningowych.

Porównanie przeszukania przestrzeni dwóch hiperparametrów (x_1, X_2) z wykorzystaniem obu opisywanych metod przedstawione zostało na wykresie 4.9. Optymalne wartości parametrów znajdują się na skrzyżowaniu zacienionych pól.

4.2.3 Ewaluacja modelu

Proces ewaluacji modelu polega na wyznaczeniu zdefiniowanych wskaźników jakości modelu zarówno dla danych treningowych, jak i testowych. Na podstawie uzyskanych wyników oceniana jest użyteczność modelu oraz podejmowana jest decyzja czy stworzony model może



Rysunek 4.9: Porównanie przeszukania przestrzeni hiperparametrów z wykorzystaniem metody siatki oraz metody losowej.

zostać zaimplementowany w systemie, czy konieczne jest ponowne przeprowadzenie procesu modelowania modelu, aktualizując proces o uzyskaną wiedzę, np. konieczność dodania nowych cech, konieczność zebrania większej próbki danych. Wartości metryk pozwalają również na ocenę różnych klas modeli.

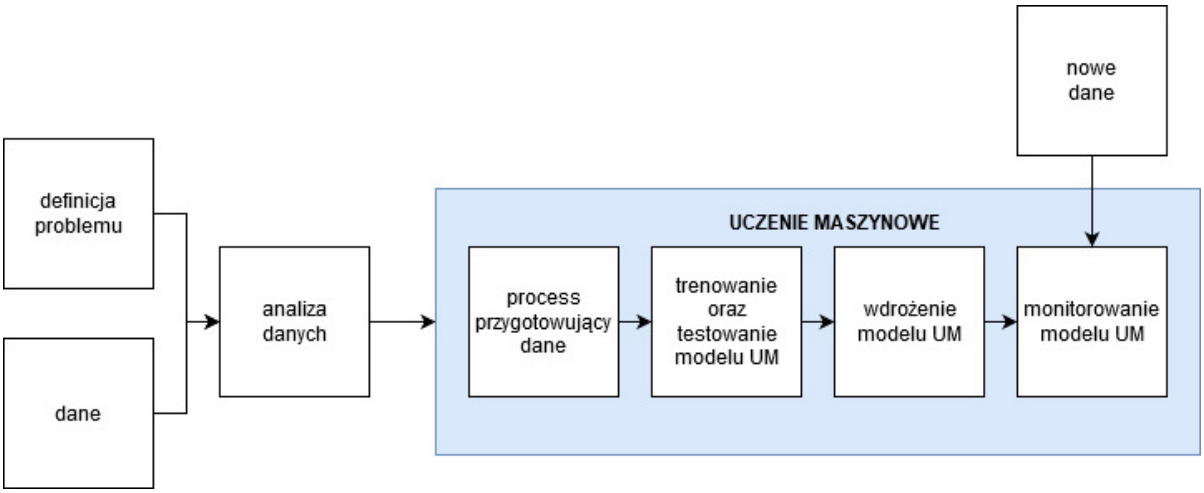
4.3 Implementacja

Celem budowy modelu UM jest rozwiązywanie postawionych problemów, natomiast modele UM mogą spełnić swoje zadanie jedynie, wtedy gdy został wdrożony i jest aktywnie wykorzystywany. W związku z tym wdrażanie modelu jest tak samo ważne, jak budowanie modelu. Stworzenie modelu oraz jego wdrożenie stanowi początek cyklu życia modelu, którego podstawowa forma przedstawiona została na wykresie 4.10.

Pojęcie modelu UM z punktu widzenia systemu zdolengo do generowania prognoz na podstawie nowych danych jest definiowane jako:

Definicja 1. Model UM - połączenie algorytmu oraz danych konfiguracyjnych pozwalających wyznaczyć predykcje na podstawie nowego zestawu danych.

Algorytm w modelu UM może być złożoną strukturą, taką jak sieci neuronowe, natomiast dane konfiguracyjne stanowią w tym przypadku wartości wag połączeń między neuronami. Działanie modelu przedstawiono na wykresie 4.11;model otrzymuje nowe dane i na ich



Rysunek 4.10: Cykl życia modelu UM.

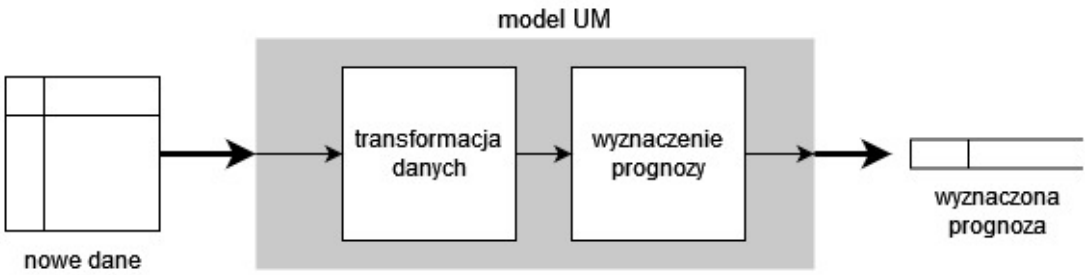
podstawie wyznacza predykcję. Dane przekształcane są do postaci wykorzystywanej przez model, a z wykorzystaniem wyznaczonych zależności wyznaczana jest prognoza. Stworzony model UM może być wykorzystywany dwojako:

- jako narzędzie wykorzystywane offline (ang. batch production) – predykcje dla nowych danych wyznaczane są na żądanie po uprzednim przygotowaniu danych. Częstotliwość wyznaczanych prognoz jest nieregularna. Dane zbierane są z opóźnieniem, np. po uprzednim manualnym zebraniu danych z wszystkich liczników;
- jako narzędzie wykorzystywane online (ang. real-time production) – predykcje wyznaczane są automatycznie, często cyklicznie na podstawie danych zbieranych w czasie rzeczywistym lub około rzeczywistym.

Modele offline, które wymagają niewielkich nakładów inżynieryjnych, są pomocne w wizualizacji, planowaniu i prognozowaniu przy podejmowaniu decyzji biznesowych. Z drugiej strony modele online wymagają znacznego wysiłku inżynieryjnego i są używane do wspomagania bieżącej pracy oraz rekomendacji.

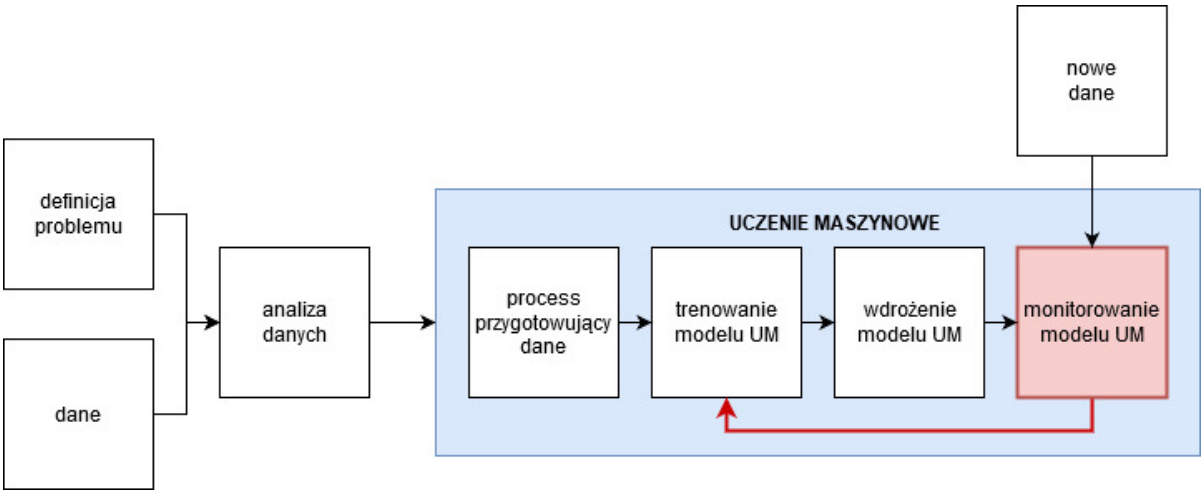
Monitorowanie jakości predykcji modelu

Po wdrożeniu modelu do produkcji konieczne jest monitorowanie, jak dobrze działa model. Pod uwagę należy wziąć kilka obszarów wydajności modelu, gdzie każdy z nich będzie posiadał własne metryki i pomiary wpływające na cykl życia modelu.



Rysunek 4.11: Model uczenia maszynowego.

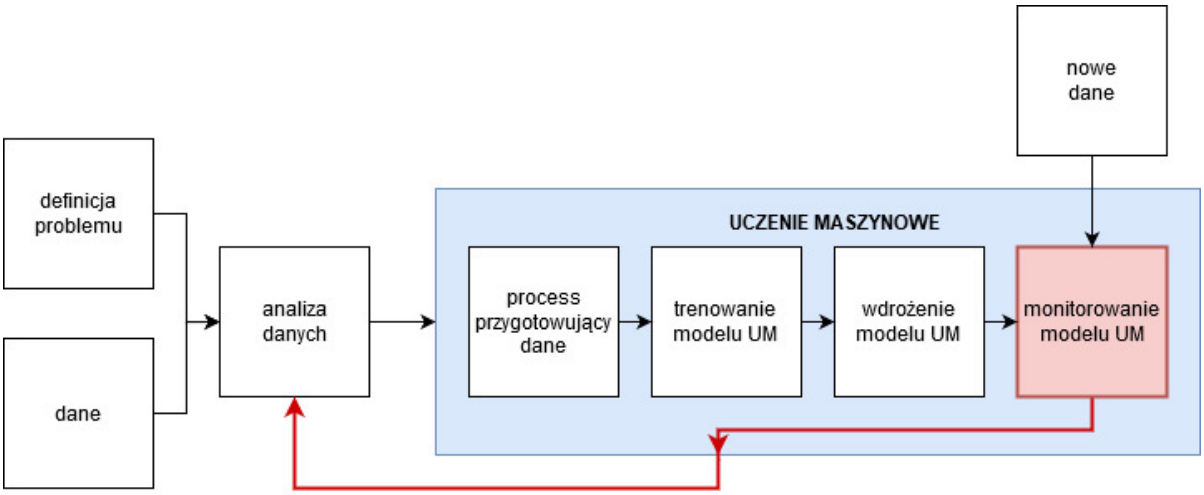
Dryf modelu (ang. model drift) jest terminem stosowanym do przedstawiania zmian w zdolnościach predykcyjnych modelu, tj. dokładności prognoz generowanych przez model. Model UM wdrożony w dynamicznie zmieniającym się systemie, w którym dane są regularnie zbierane oraz aktualizowane dostępne zestawy danych może znacząco się zmieniać w krótkich okresach czasu. W związku z powyższym wartości cech wykorzystywane do wyznaczania prognoz również mogą się dynamicznie zmieniać. W takim przypadku, gdy jakość predykcji modelu pogorszy się, konieczne jest ponowne trenowanie modelu z wykorzystaniem danych rozszerzonych o zestaw danych zebranych od ostatniego wdrożenia. Proces ten przedstawiony został na wykresie 4.12.



Rysunek 4.12: Cykl życia modelu: dryf modelu.

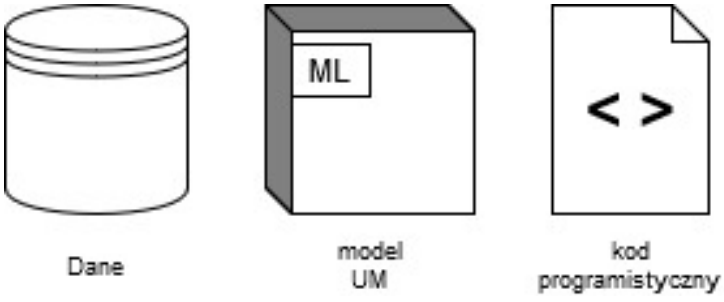
W pewnych przypadkach ponowne trenowanie modelu z wykorzystaniem rozszerzonego zestawu danych może nie przynieść pożądanych rezultatów. Są to przypadki, gdy dane wejściowe zmieniły się w czasie na tyle mocno, że wyselekcjonowane cechy w procesie selekcji cech nie stanowią obecnie wystarczająco dobrego zestawu danych pozwalającego na mapowanie zależności między zmiennymi wejściowymi oraz cechą prognozowaną. W takim

przypadku konieczne jest ponowne wykonanie kroku analizy danych i sprawdzenie całego procesu tworzenia modelu łącznie z etapem inżynierii cech. Proces ten może przykładowo wymagać dodania nowych źródeł danych lub zmiany typu, lub architektury modelu, np. pogłębienia sieci neuronowej. Taki przypadek nazywany jest dryfem danych i przedstawiony został na wykresie 4.13.



Rysunek 4.13: Cykl życia modelu: dryf danych.

W cyklu życia modelu konieczne jest stworzenie procesu odpowiedzialnego za mierzenie jakości predykcji modelu UM wyznaczanego na postawie nowo dostępnych danych. W przypadku pogorszenia jakości poniżej akceptowalnego poziomu rozpoczynany jest nowy proces, który inicjuje ponowne trenowanie modelu UM, a następnie wdraża najnowszy model. Ponowne trenowanie modelu wykorzystuje również nowe dane zabrane od czasu ostatniego wdrożenia modelu. Autorzy [66] podkreślają, że we wdrożonych systemach ML, jedynie małą część systemu stanowi kod programistyczny UM, gdyż konieczna jest szeroka infrastruktura informatyczna. Kod programistyczny, model UM oraz dane wykorzystane do jego budowy stanowią obszary możliwych zmian w procesie wdrażania systemu UM.



Rysunek 4.14: Obszary zmienności systemu uczenia maszynowego.

Podsumowując, cykl życia modelu UM zaczyna się po jego wdrożeniu, tzn. wykorzystywania zgodnie ze zdefiniowanym na wstępie modelem. Po wdrożeniu modelu w wyniku pojawiania się nowych danych oraz zmienności systemu, który stanowi obszar prognozy modelu może nastąpić pogorszenie jakości prognoz modelu. Obszary zmienności modelu UM obejmują trzy obszary przedstawione na wykresie 4.14, tj. obszar danych, obszar modelu oraz obszar kodu.

Rozdział 5

Warszawska sieć ciepłownicza

Warszawski system ciepłowniczy (WSC) jest największym systemem ciepłowniczym w kraju i jednym z największych w Europie – całkowita długość sieci ciepłowniczej wynosi 1700 km sieci (ok. 300 km magistrali oraz po ok. 700 km rurociągów rozdzielczych i przyłączy) i zasila 190 km^2 obszaru miejskiego. Schemat WSC przedstawony jest na wykresie 5.1. WSC dostarcza ciepło do ponad 19 tysięcy obiektów na terenie Warszawy, pokrywając ok. 80 proc. potrzeb cieplnych miasta, które liczy ponad 1,7 miliona mieszkańców. Sieć dostarcza ciepło na potrzeby ciepłej wody użytkowej oraz ogrzewania do dużej grupy obiektów takich jak: budynki mieszkalne, szkoły, biurowce, centra handlowe, stadiony, itp., a wielkość rocznej sprzedaży ciepła z tej sieci wynosi ok. 30 do 36 PJ w zależności od intensywności sezonu grzewczego. Rocznie za pośrednictwem sieci ciepłowniczej dostarczane jest odbiorcom ok. 10000 GWh ciepła. Maksymalne zapotrzebowanie WSC na energię cieplną to ok 3600 MWt, a minimalne w ciągu lata wynosi ok. 290–300 MWt. Poziom strat ciepła waha się pomiędzy 10 a 12 proc., przy czym ok. 1,5 proc. strat wynika z ubytków wody sieciowej [2].

Elementami systemu ciepłowniczego pośredniczącymi w przesyle wody sieciowej do odbiorców są cztery przepompownie (Batorego, Golędzinów, Marymont, Żwirki), które pracują w razie zbyt niskiego ciśnienia oraz przepływu. W skład przepompowni wchodzi armatura, układ sterujący oraz pompy pompujące wodę w rurociągach zasilanych z elektrociepłowni oraz te pompujące wodę powrotną. Ważnym elementem systemu ciepłowniczego jest węzeł cieplny, czyli zakończenie sieci ciepłowniczej w budynku. W Warszawie jest około 17 000 węzłów składających się z jednostopniowego wymiennika centralnego ogrzewania oraz dwustopniowego wymiennika ciepłej wody. Każdy węzeł wyposażony jest w licznik ciepła oraz automatykę pogodową, czyli sterownik elektroniczny wykorzystujący czujki temperatury

powietrza, wody zasilającej i powrotnej do sterowania ilością wymienianego ciepła. Urządzenie

Rysunek 5.1: Schemat warszawskiej sieci ciepłowniczej [1].

to reaguje na temperaturę zewnętrzną i w ten sposób dopasowuje ilość ciepła dostarczanego do budynku. Należy również wspomnieć o komorach ciepłowniczych, których na terenie Warszawy znajduje się około 5000. Wyróżnikiem warszawskiej sieci ciepłowniczej jest budowa rozgałęźno-pierścieniowa – dzięki niej każdy obszar może być zasilany z różnych źródeł, a odbiorcy mają pewność dostaw w razie awarii jednego z nich.

Źródła ciepła w warszawskim systemie ciepłowniczym to:

- dwie elektrociepłownie – Siekierki i Żerań – pracujące przez cały rok, produkujące w procesie skojarzonym (tzw. kogeneracja) ciepło oraz energię elektryczną;
- dwie ciepłownie – Kawęczyn i Wola – będące źródłami szczytowymi uruchamianymi wyłącznie w okresach największego zapotrzebowania klientów na energię cieplną;
- niewielka spalarnia odpadów komunalnych wytwarzająca energię elektryczną i cieplną, należąca do MPO.

Źródła te dysponują łączną mocą cieplną 4635 MWt i mocą elektryczną 1006 MWe.

Rozdział 6

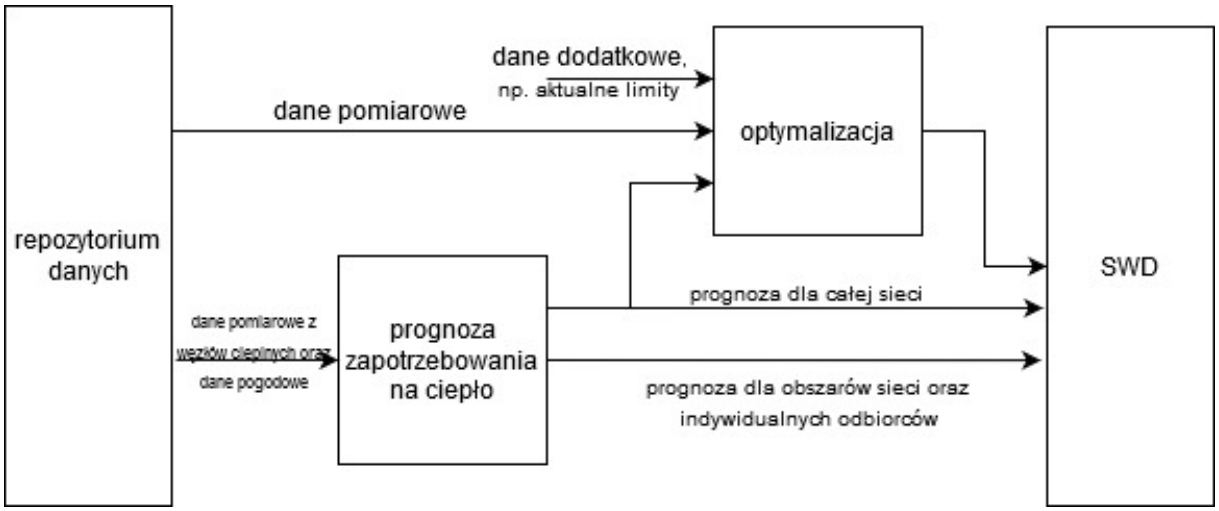
System prognozowania zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej

Stworzony przez autorkę we współpracy z firmą Transition Technologies S.A. SPROG stanowi element Systemu Wsparcia Decyzji (SWD), którego podstawowym celem jest optymalizacja pracy w warszawskiej sieci ciepłowniczej zarządzanej przez Veolia Energia Warszawa S.A. Optymalizacja pracy sieci jest ukierunkowana na obniżenie kosztów dystrybucji ciepła przez definiowanie optymalnych parametrów pracy źródeł ciepła pozwalających na zapewnienie komfortu cieplnego wszystkich użytkowników sieci przy minimalnych stratach ciepła w sieci. System wsparcia decyzji na założonym horyzoncie czasowym (domyślnie 48 lub 120 godzin) dla każdej godziny horyzontu ustala optymalne wartości:

- stanów pracy każdego źródła ciepła (decyduje o włączeniu bądź zatrzymaniu pracy źródła ciepła),
- temperatury wody i ciśnienia wody na wyjściu każdego źródła ciepła,
- ciśnienia podniesienia na przepompowniach, zarówno na zasilaniu, jak i na powrocie.

Optymalne wartości powyższych parametrów dobierane są tak, aby zapewnić dostawę ciepła do wyróżnionych węzłów i komór krytycznych oraz spełnienie ograniczeń technicznych, operacyjnych i kontraktowych związanych z pracą źródeł, czyli przepompowni, komór i węzłów.

Optymalne wartości parametrów sieci dla domyślnego horyzontu 48 oraz 120 godzin



Rysunek 6.1: Ogólna zasada działania SWD.

wyznaczane są codziennie na podstawie zbieranych danych pomiarowych (wczesne godziny poranne) tak, aby stanowić pomoc w pracy operatorów sieci.

Uruchomienie algorytmu optymalizacji wyzwala sekwencję kroków przedstawionych na wykresie 6.1, a historyczne oraz aktualne dane pomiarowe przesyłane są z repozytorium danych do SPROG, gdzie wyznaczana jest prognoza zapotrzebowania na ciepło dla zadanego horyzontu predykcji. Następnie algorytm optymalizacji, wykorzystując historyczne oraz aktualne dane pomiarowe wraz z wynikiem prognozy zapotrzebowania na ciepło dla całej sieci oraz danymi dodatkowymi, takimi jak ograniczenia przestrzeni decyzyjnej rozwiązuje zadanie optymalizacji. Wyniki SPROG oraz optymalizacji przesyłane są do SWD, gdzie prezentowane są użytkownikom systemu, tj. operatorom sieci oraz analitykom.

Z uwagi na fakt, iż obszarem optymalizacji są straty w wyniku przesyłu ciepła, optymalizator pracy sieci jako sygnał wejściowy przyjmuje zapotrzebowanie na ciepło rozumiane jako suma konsumpcji ciepła przez odbiorców (formuła 3.2).

6.1 Definicja założeń systemu

Zgodnie z metodologią budowania systemów opartych na technikach UM przedstawioną w rozdziale 4przed przystąpieniem do tworzenia systemu wymagane jest zdefiniowanie założeń systemu. Dla definicji systemu SPROG założono:

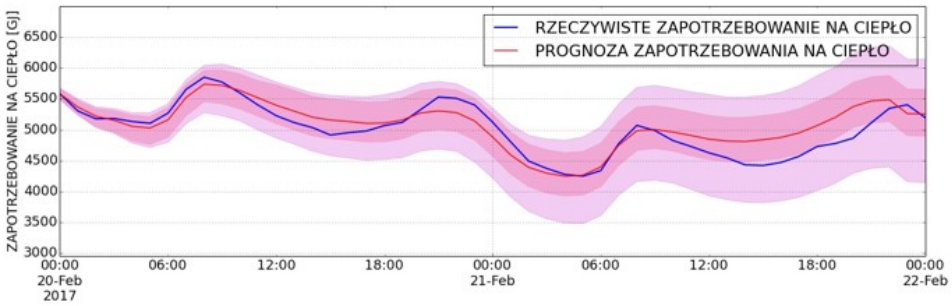
- podstawowym celem SPROG jest pomoc w realizacji prognoz zapotrzebowania na ciepło. System będzie wyznaczał następujące prognozy:

  - prognoza zapotrzebowania na ciepło dla całej sieci;
  - prognozy zapotrzebowania na ciepło dla zdefiniowanych obszarów sieci;

- realizacja prognoz możliwa będzie dzięki stworzeniu modeli prognostycznych opartych na technikach uczenia maszynowego. Dla każdej wymaganej prognozy stworzony zostanie oddzielny model prognostyczny;
- modele prognostyczne stworzone zostaną z wykorzystaniem uczenia z nauczycielem [36];
- wartości zapotrzebowania na ciepło będą wyznaczane przez system w oparciu o dane pomiarowe z węzłów. System będzie wyznaczał:
  - poziom rzeczywistego zapotrzebowania na ciepło dla całej sieci;
  - – poziom rzeczywistego zapotrzebowania na ciepło dla zdefiniowanych obszarów sieci;
- wymagane dane obejmują dane pomiarowe z węzłów cieplnych zawierające pomiar zużycia ciepła w węźle oraz dane pogodowe zawierające co najmniej pomiar temperatury zewnętrznej;
- minimalny zakres wymaganych danych obejmuje dwa pełne lata kalendarzowe;
- jakość modeli prognostycznych oceniana będzie na podstawie wartości współczynników dopasowania modeli MAPE oraz R^2;
- jakość modeli prognostycznych oceniana będzie z podziałem na sezony: grzewczy, letni oraz przejściowy;
- dodatkowym wynikiem modeli prognostycznych, dla każdej wartości prognozowanej, będą wartości określające przedziały ufności. Przykład prognozy dla całej sieci wraz z rzeczywistymi wartościami przedstawiony został na wykresie 6.2. Historyczne wartości zapotrzebowania na ciepło zaznaczone są kolorem niebieskim, natomiast wynik prognozy zawiera:
  - niepewność wyniku uwzględniającą niedoskonałość modelu prognostycznego, obszar oznaczony kolorem jasno czerwonym;

– niepewność wyniku uwzględniającą dodatkowo niepewność prognozy pogody, obszar oznaczony kolorem czerwonym;

- dokładność 24-godzinnych predykcji będzie równa lub lepsza od wyników prezentowanych w literaturze naukowej, tj. MAPE dla sezonu grzewczego nie przekroczy 4,76 proc.;
- jakość predykcji modeli będzie cyklicznie oceniania i w przypadku pogorszenia predykcji powyżej poziomu krytycznego nastąpi autokalibracja modeli na podstawie danych historycznych uzupełnionych o dane zebrane od momentu ostatniej rekalibracji modeli.

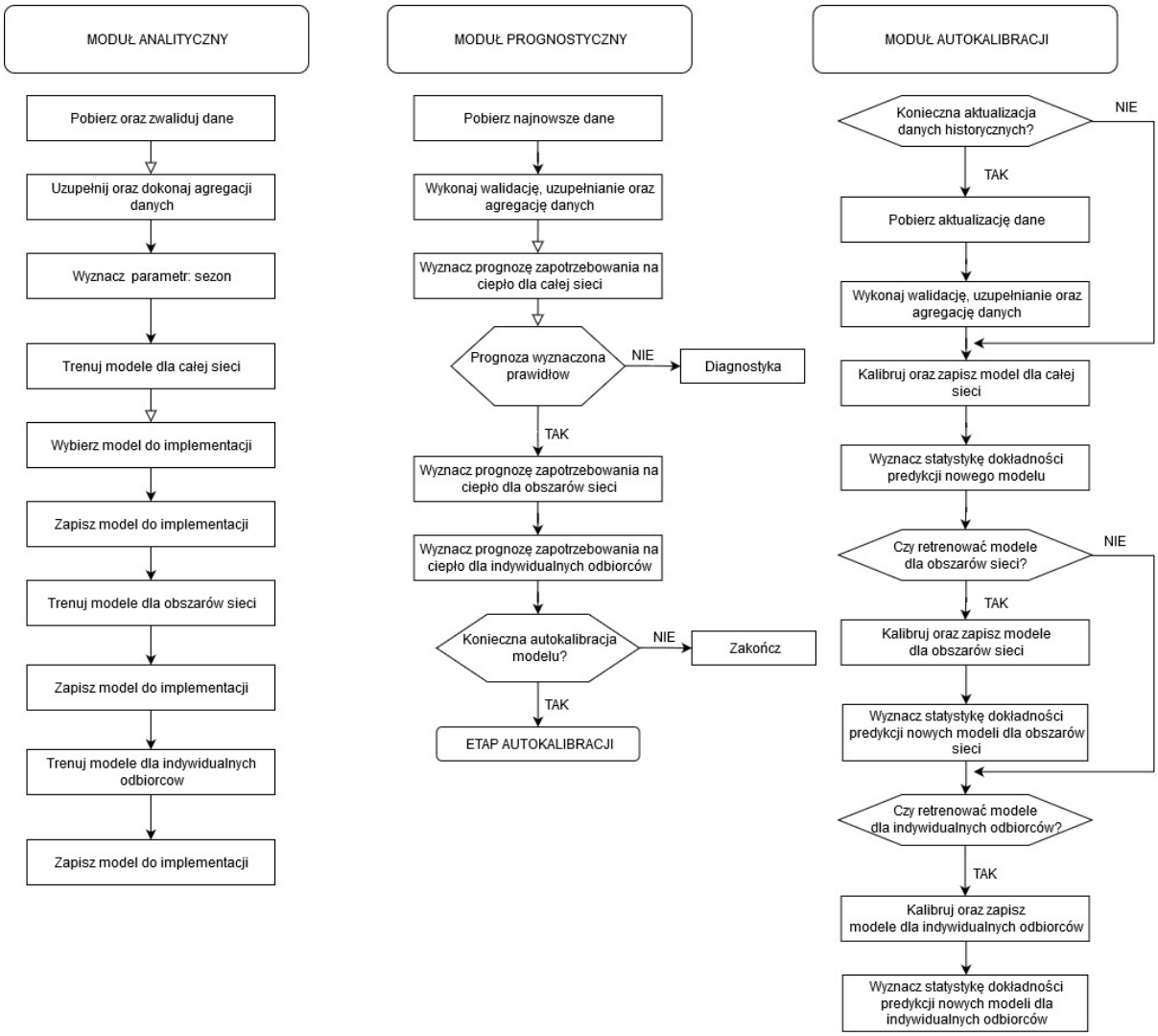


Rysunek 6.2: Wygląd prognozy dla całej sieci w systemie SWD.

Tworzenie systemu poprzedzone jest etapem analitycznym, następnie tworzony jest sam system, który bazując na wypracowanych algorytmach oraz modelach wyznacza prognozy zapotrzebowania na ciepło. Podstawowe algorytmy modułów oraz kolejność ich wykonania przedstawione zostały na rysunku 6.3.

6.2 Etap analityczny

Etap analityczny, szczegółowo przedstawiony w niniejszej rozprawie, koncentruje się na stworzeniu modeli prognostycznych będących podstawą funkcjonalności tworzonego systemu. Wszystkie rozwiązania opracowywane w ramach etapu analitycznego są adaptowane w systemie, tj. metodologia walidacji danych oraz uzupełniania brakujących danych. Typ i klasa modeli UM wpływa na regułę prognostyczną determinującą funkcjonalność modułu prognostycznego. Komunikacja między modułami odbywa się przy pomocy specjalnych



Rysunek 6.3: Główne moduły systemu oraz ich algorytmy.

zmiennych znajdujących się w bazie danych lub plikach płaskich.

6.2.1 Dane surowe

System prognostyczny prezentowany w niniejszej rozprawie realizuje wskazane cele przy założeniu, że dostępne są wymagane dane surowe. W celu poprawy jakości modeli prognostycznych możliwe jest zapewnienie dodatkowych danych, jednak nie jest to czynnik krytyczny decydujący o funkcjonalności systemu.

Krytyczne dane surowe, tj. dane wymagane do stworzenia modeli prognostycznych, oraz dodatkowe dane pozwalające na potencjalne zwiększenie dokładności predykcji modeli

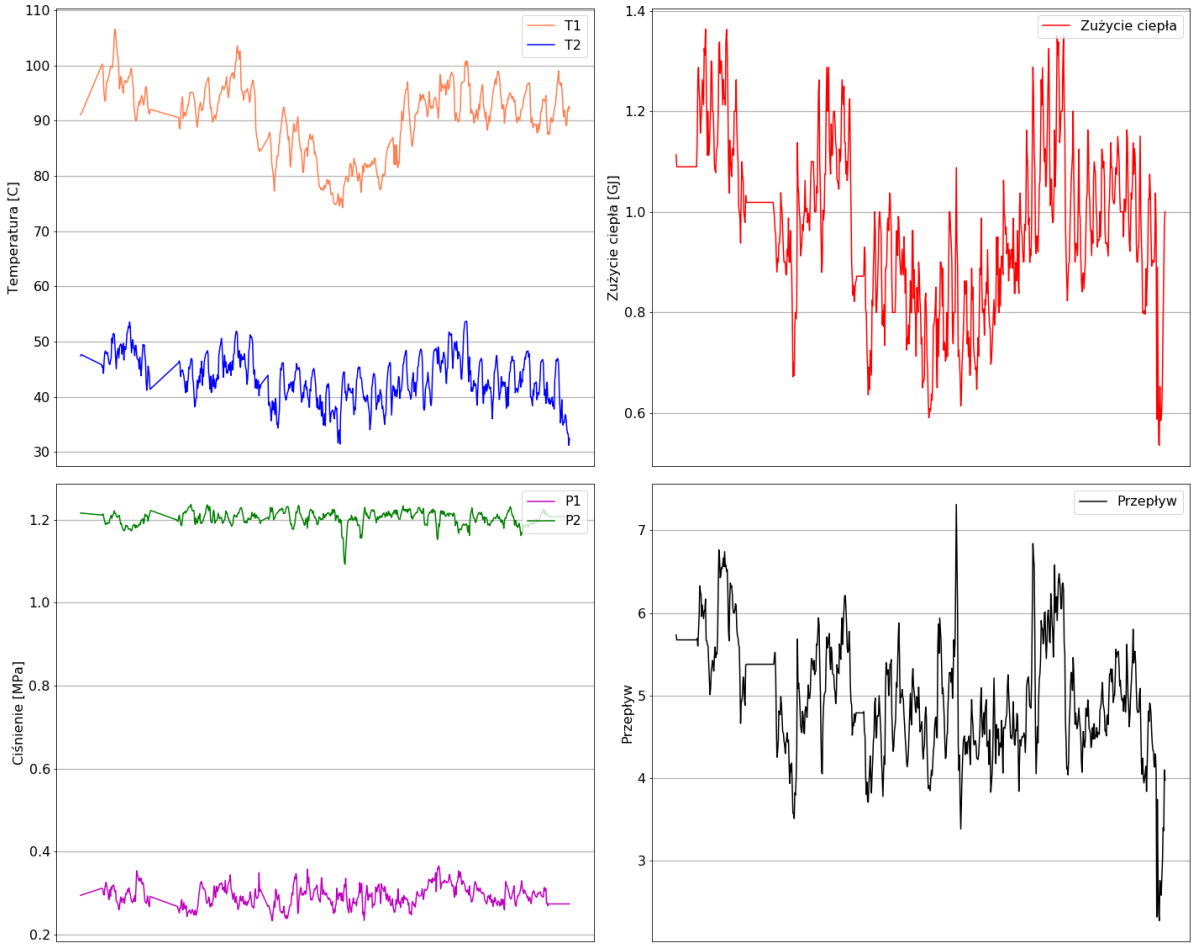
Tabela 6.1: Wymagane dane surowe.

Typ Danych	Klasa danych	Nazwa
Dane pomiarowe z węzłów cieplnych	krytyczne	Pomiar zużycia ciepła
	dodatkowe	Pomiar temperatury zasilającej węzeł
		Pomiar temperatury powrotnej z węzła
		Pomiar ciśnienia zasilającego węzeł
		Pomiar ciśnienia na wyjściu z węzła
		Pomiar przepływu wody w węźle
Dane pogodowe	krytyczne	Temperatura zewnętrzna
	dodatkowe	Prędkość wiatru
		Kierunek wiatru
		Wilgotność
		Nasłonecznienie
		Zachmurzenie
		Inne
Dane dodatkowe dotyczące węzłów cieplnych	krytyczne - możliwe do oszacowania	Temperatura włączenia centralnego ogrzewania
	dodatkowe	moc zamówiona
		typ budynku

wykorzystane podczas tworzenia oraz testów systemu przedstawione zostały w tabeli 6.1.

Wykres prezentujący wybrany miesiąc w sezonie grzewczym dla danych pomiarowych z wymiennika ciepła przedstawiony został na rysunku 6.4.

Kolejnym krytycznym czynnikiem rzutującym na dokładność modeli prognostycznych jest długość horyzontu danych historycznych. W celu efektywnego trenowania modeli oraz poprawnej oceny ich jakości wymagany jest minimum dwuletni horyzont danych historycznych. Dwuletni horyzont tego typu danych pozwala na stworzenie dwóch zbiorów danych – zbiór danych treningowych oraz zbiór danych testowych, które zawierają reprezentatywny okres zmienności danych. Należy tutaj zaznaczyć, że w przypadku większości nieliniowych empirycznych modeli prognostycznych wraz z wzrostem długości horyzontu danych treningowych rośnie dokładność predykcji modelu. W świetle prognozowania zapotrzebowania na ciepło istotne jest, aby zbiór danych treningowych zawierał możliwie najwięcej okresów przejściowych między sezonem letnim oraz grzewczym.



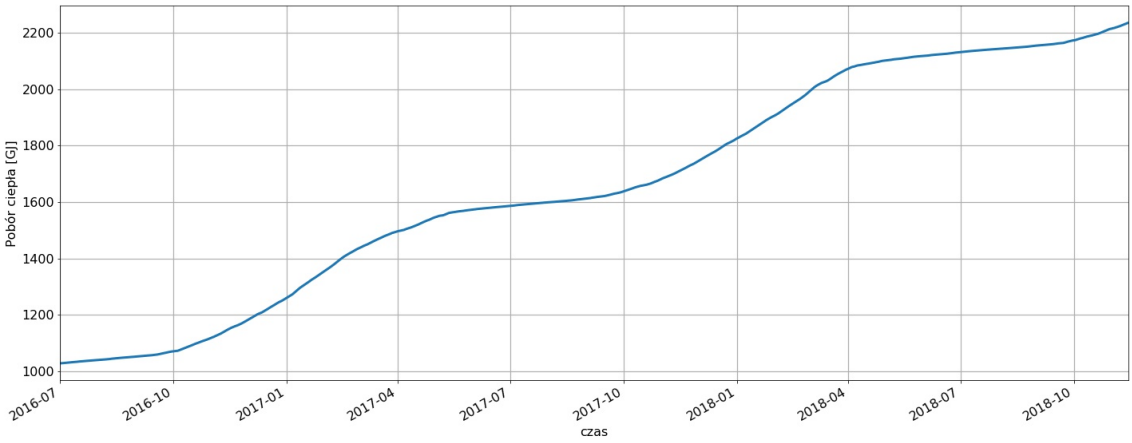
Rysunek 6.4: Przebieg przykładowych danych pomiarowych z węzła cieplnego.

Do stworzenia prezentowanych w niniejszej pracy modeli prognostycznych wykorzystane dane pomiarowe z okresu lipiec 2015 – lipiec 2018.

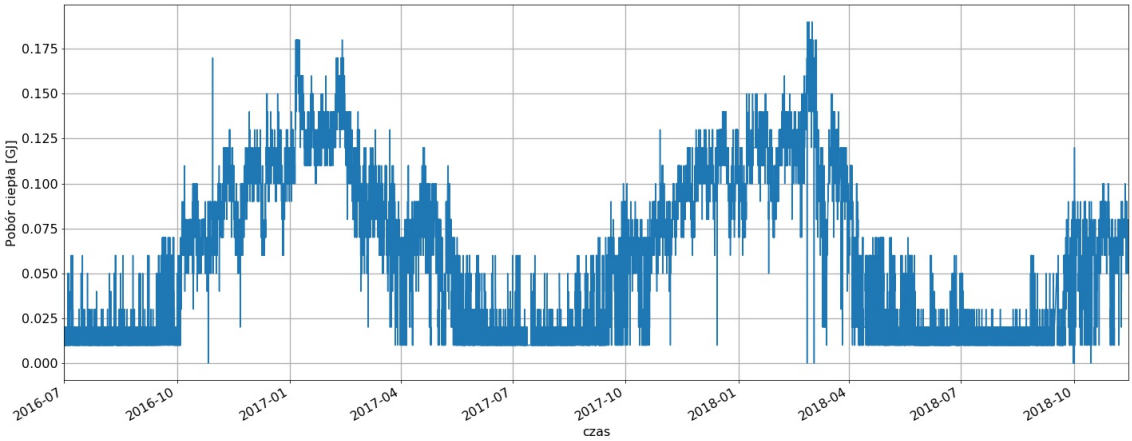
6.2.2 Analiza eksploracyjna oraz walidacja danych z węzłów cieplnych

Zebrane dane surowe należy przeanalizować pod kontem ich kompletności oraz jakości. Następnie konieczna jest identyfikacja danych odstających, tzn. walidacja danych.

Dane z węzłów cieplnych zawierały pomiary zużycia ciepła z 15 600 węzłów cieplnych. Pomiar zużycia ciepła, rejestrowany z godzinnym krokiem czasowym, jest pomiarem pośrednim wyznaczanym na podstawie pomiarów przepływu wody i temperatury na wejściu oraz powrocie z węzła. Pomiar rejestruje ilość energii cieplnej dostarczonej do obiektu w danym okresie rozliczeniowym, wykres pomiaru zużycia ciepła dla przykładowego węzła cieplnego przedstawiony jest na wykresie 6.5. W celu wyznaczenia zapotrzebowania na ciepło



Rysunek 6.5: Pomiar zużycia ciepła dla przykładowego węzła cieplnego.

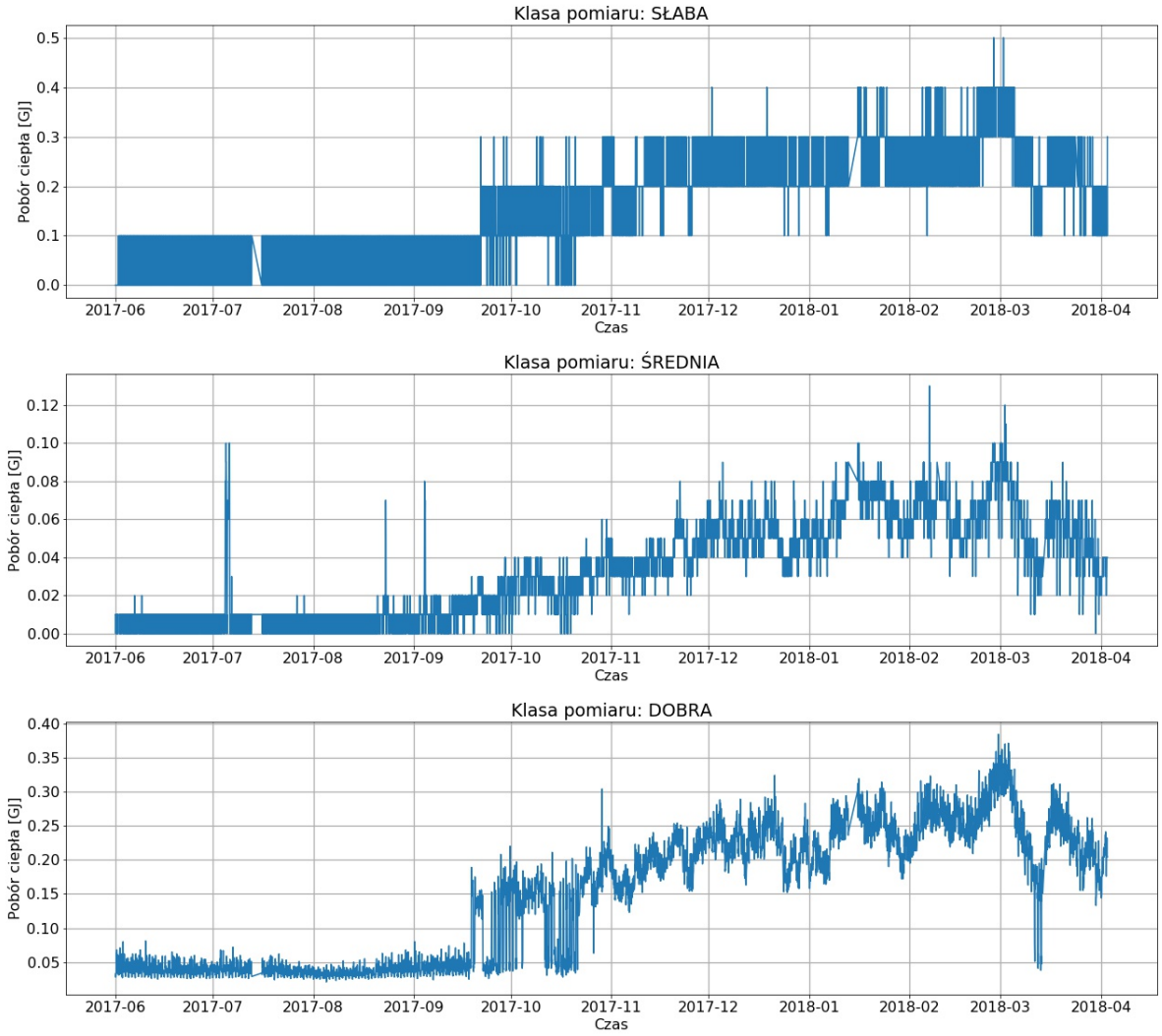


Rysunek 6.6: Zapotrzebowanie na ciepło w ujęciu godzonowym dla przykładowego węzła cieplnego.

w ujęciu godzinowym wyznaczana jest różnica między wartościami pomiarów dla godziny t oraz godziny $t-1$. W dalszej części pracy termin zapotrzebowanie na ciepło rozumiany jest jako zużycie ciepła w ujęciu godzinowym. Wykres wyznaczonego zużycia ciepła w ujęciu godzinowym dla przedstawianego na wykresie 6.5 wymiennika przedstawiony jest na wykresie 6.6.

Wizualna analiza danych pomiarowych zapotrzebowania na ciepło wykazała, że pomiary zużycia ciepła mają różną czułość oraz różne poziomy zmian wartości zapotrzebowania na ciepło. Pomiary podzielono na trzy klasy pomiarów w zależności od liczby zapisywanych unikalnych wartości:

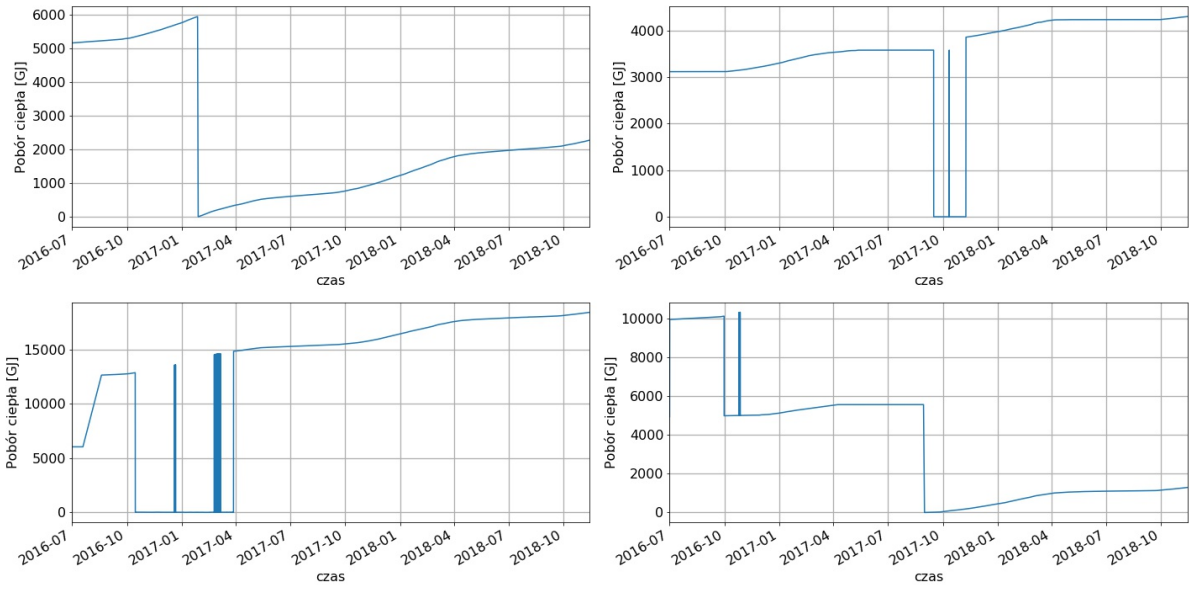
- pomiary klasy SŁABEJ, które przyjmują maksymalnie trzy unikalne wartości (53 proc. pomiarów);



Rysunek 6.7: Zapotrzebowanie na ciepło w ujęciu godzonowym dla klas pomiaru: SŁABA, ŚREDNIA i DOBRA.

- pomiary klasy ŚREDNIEJ, które przyjmują od 3 do 5 unikalnych wartości (21 proc. pomiarów);
- pomiary klasy DOBREJ, które przyjmują powyżej pięciu unikalnych wartości (26 proc. pomiarów).

Wykresy pomiarów dla przykładowych ciepłomierzy z każdej klasy przedsatwione zostały na rysunku 6.7. Stosunek klas poszczególnych pomiarów wskazuje na zdecydowaną przewagę pomiarów klasy SŁABEJ, co daje pierwszą informację, że poziom błędów modelu zapotrzebowania na ciepło może być zwiększony z uwagi na niepełne wychwytywanie dla tych węzłów dynamicznych momentów zmian konsumpcji ciepła. Jednocześnie można przypuszczać, że wspomniane niedoskonałości pomiarów konsumpcji ciepła podczas



Rysunek 6.8: Pomiary z występującymi anomaliami dal wybranych ciepłomierzy.

wyznaczania zapotrzebowania na ciepło dla całej sieci zostaną skompensowane.

Dalsza analiza danych z ciepłomierzy pozwoliła zidentyfikować najczęściej występujące anomalie i awarie w analizowanych miernikach, które występują w danych. Zidentyfikowano następujące typy anomalii:

- wymiana ciepłomierza i związane z nią nieprawidłowe odczyty w szczególnosci znaczne skoki wartości zużycia ciepła;
- awaria pracy wymiennika lub odcięcie węzła (także planowe) obrazujące się w braku zużycia ciepła;
- anomalie i błędy czujników pomiarowych, np. wysunięcie czujnika temperatury obrazujące się w nieprawidłowych wartościach danych pomiarowych;
- punktowe atypowo wysokie wartości danych pomiarowych.

Na wykresie 6.8 przedstawione zostały wykresy dla wybranych ciepłomierzy, dla których wystąpiły anomalie w zapisie danych opisane wyżej. Zauważalne jest, że po wystąpieniu pierwszej anomalii, która może być wymianą ciepłomierza pojawiają się w konsekwencji kolejne anomalie. Dlatego dla danych z ciepłomierzy, w których zidentyfikowano takie anomalie wykonane zostały następujące czynności:

- dla każdego przypadku, dla którego okres z anormalnym zachowaniem przekraczał 50 proc. czasu pracy ciepłomierza, dany ciepłomierz zostawał usuwany ze zbioru danych podlegających analizie i predykcji;
- dla każdego ciepłomierza po zidentyfikowaniu wymiany ciepłomierza i/lub anormalnego zachowania analogicznego do wymiany ciepłomierza usuwany był dany okres oraz trzy dni przed wystąpieniem danego stanu oraz trzy dni po jego zakończeniu.

Tabela 6.2: Reguły walidacji pełnej.

REGUŁA	**OPIS**
Usuń wartości dla których odstęp pomiędzy pomiarami jest większy niż 1h.	Ponieważ rejestr zużytego ciepła jest wartością narastającą a wartość modelowana wyznaczana jest jako różnica wartości rejestru między odczytami (zapis co godzinę) odrzucane są pomiary dla których odstęp przekracza 1h – wartości są większe.
Usuń 12 h przed i 12h po wymianie ciepłomierza (W przypadku modelowania usuwane były 3 dni przed i 3 po).	Jako moment wymiany ciepłomierza traktuje się zarejestrowanie negatywnej wartości różnicy między odczytami.
Dla sezonu zimowego usuń odczyty z zerowym przepływem.	Dla sezonu zimowego zakłada się że każdy węzeł jest odbiorcą ciepła. Dla sezonu letniego i przejściowego dopuszcza się wartości zerowe.
Usuń godziny dla których różnica między temperaturą zasilnia i powrotu jest niższa niż 5°C oraz temperatura zasilania jest wyższa niż 55°C	Zakłada się że każdy węzeł jest odbiorcą ciepła. Warunek temperatury zasilania pozwala pominąć nieaktywne węzły CO (sezon letni i przejściowy)
Usuń godziny dla których ciśnienie zasilania jest większe niż ciśnienie powrotu. Usuń godziny dla których ciśnienie zasilania jest większe niż 1.25 MPa. Usuń godziny dla których różnica między ciśnieniem zasilania a ciśnieniem powrotu jest mniejsze niż 0.1 MPa	Podstawowe warunki na ciśnieni, usuwające godziny z błędnym rejestrem.

Ostatnim krokiem etapu walidacji danych było stworzenie algorytmu walidującego. Z uwagi na możliwie zmienną dostępność danych pomiarowych stworzone zostały dwie funkcje walidujące, które dokonują walidacji pełnej lub uproszczonej, i których definicje przedstawiono poniżej:

Definicja 2. Walidacja pełna - walidacja danych surowych oparta na funkcji walidującej wykorzystującej dane pomiarowe z węzłów cieplnych zawierająca pomiary:

- temperatury zasilania i powrotu,
- ciśnienia zasilania i powrotu
- przepływu,
- poboru ciepła.

Definicja 3. Walidacja uproszczona - walidacja danych surowych oparta na funkcji walidującej wykorzystującej dane pomiarowe z węźłów cieplnych zawierające pomiary:

- poboru ciepła.

Reguły walidujące dla walidacji pełnej przedstawione zostały w tabeli 6.2, natomiast reguły walidujące dla walidacji uproszczonej przedstawione zostały w tabeli 6.3.

Tabela 6.3: Reguły walidacji uproszczonej.

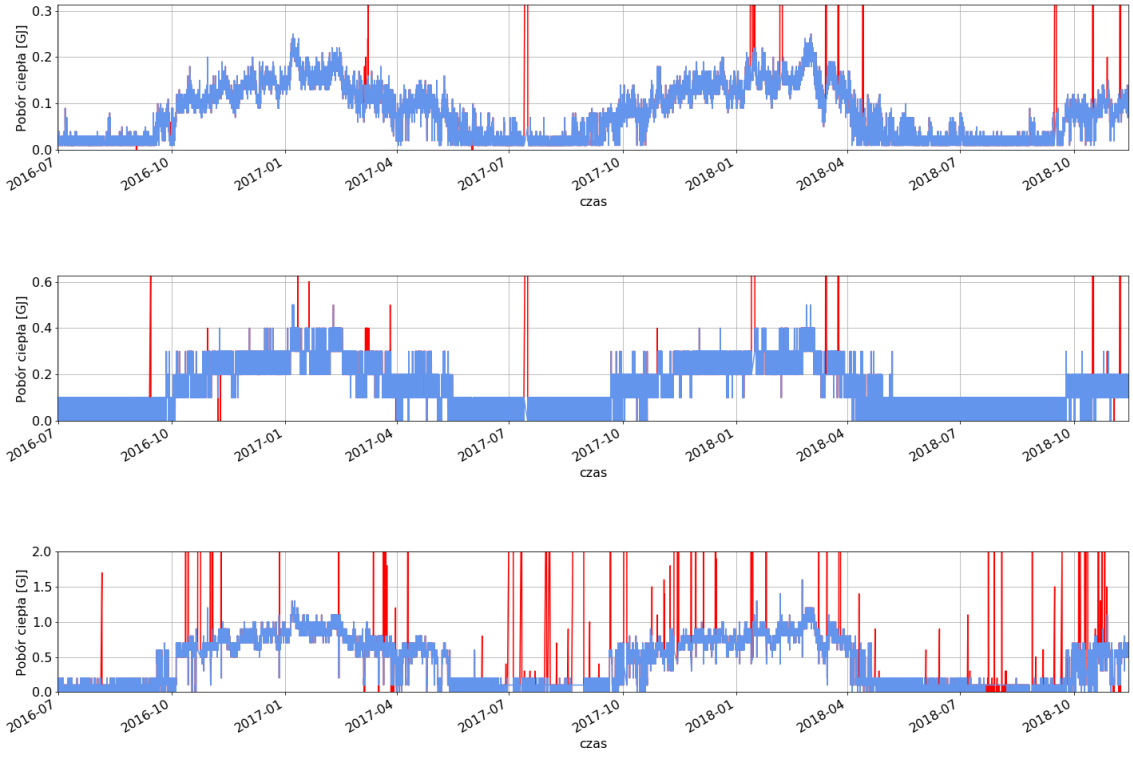
REGUŁA	OPIS
Usuń wartości dla których odstęp pomiędzy pomiarami jest większy niż 1h.	Ponieważ rejestr zużytego ciepła jest wartością narastającą a wartość modelowana wyznaczana jest jako różnica wartości rejestru między odczytami (zapis co godzinę) odrzucane są pomiary dla których odstęp przekracza 1h – wartości są większe.
Usuń 12 h przed i 12h po wymianie ciepłomierza (W przypadku modelowania usuwane były 3 dni przed i 3 po).	Jako moment wymiany ciepłomierza traktuje się zarejestrowanie negatywnej wartości różnicy między odczytami.
Dla sezonu zimowego usuń odczyty z wartością zerową.	Dla sezonu zimowego zakłada się że każdy węzeł jest odbiorcą ciepła. Dla sezonu letniego i przejściowego dopuszcza się wartości zerowe.
Usuń wartości z dużymi skokami wartości	Algorytm wyznacza pochodną zużycia ciepła (druga pochodna wartości z rejestru) i usuwa wartości większe niż kwantyl 0.95 rozkładu drugiej pochodnej dla danego sezonu.
Usuń wartości z nagłym spadkiem odbioru ciepła	Algorytm usuwa wartości niższe niż kwantyl 0.01 rozkładu zużycia ciepła dla danego sezonu.

Przykłady efektów działania algorytmu walidacji dla trzech wybranych ciepłomierzy przedstawione zostały na wykresach 6.9. Kolorem czerwonym przedstawione zostały dane

Tabela 6.4: Dane pogodowe

parametr	jednostka
temperatura zewnętrzna	°C
prędkość wiatru	m\s
kierunek wiatru	deg
wilgotność	%
nasłonecznienie	W\m^2
zachmurzenie	8 stopniowa skala

surowe, natomiast kolorem niebieskim przedstawiono dane otrzymane po procesie walidacji. Zauważalne jest działanie algorytmu, które w głównej mierze polega na usunięciu pomiarów o zbyt wysokiej wartości wynikającej np. z nieprawidłowych pomiarów czy wymiany miernika.

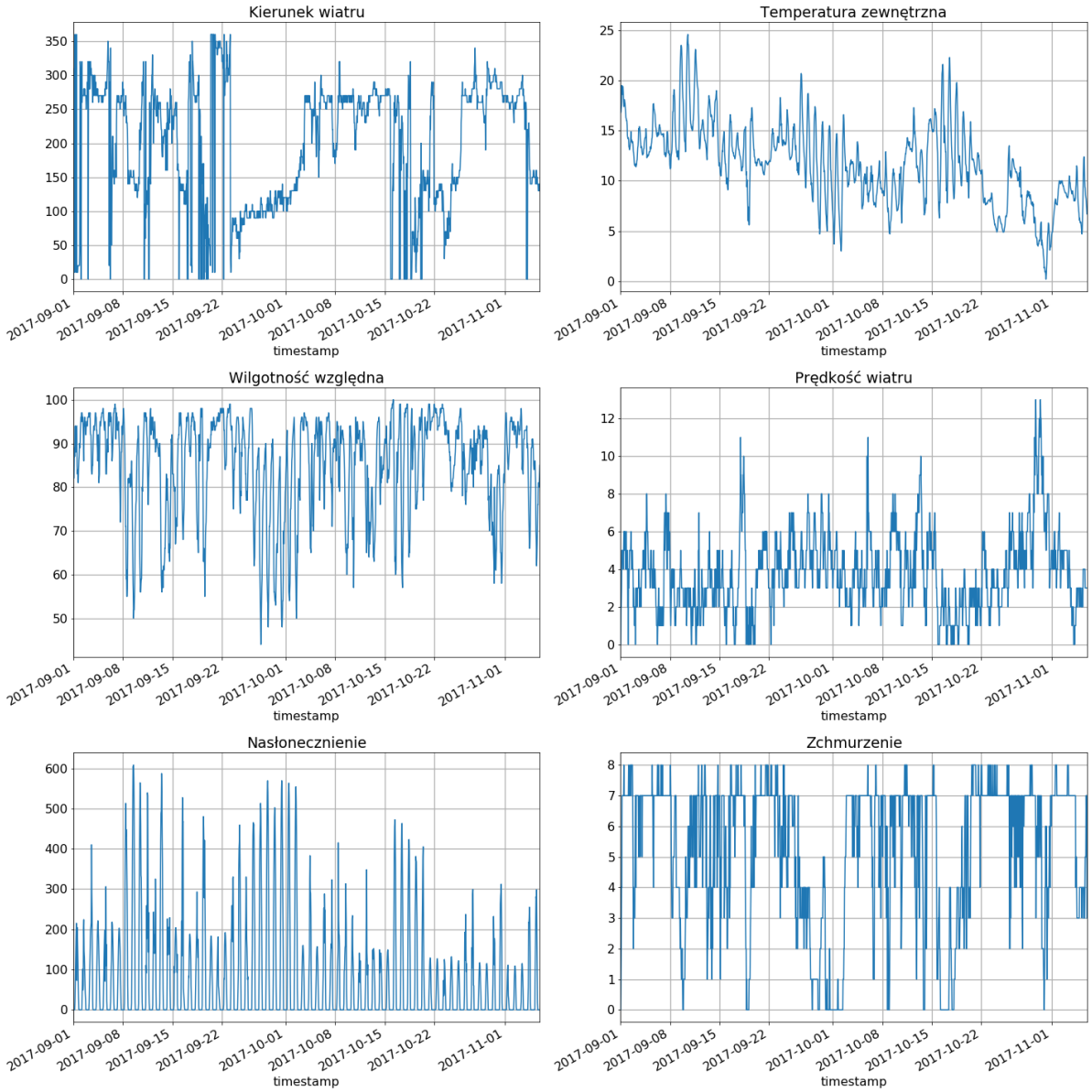


Rysunek 6.9: Przykłady działania algorytmu walidacji dla trzech różnych odbiorców.

6.2.3 Analiza eksploracyjna oraz walidacja danych pogodowych

Dane pogodowe obejmowały historyczne pomiary oraz historyczne prognozy sześciu parametrów, które przedstawione zostały w tabeli 6.4. Prognozy pogody obejmowały horyzont 48 godzin. Wymienione parametry odnosiły się do wartości dla Warszawy oraz jej dzielnic:

- Mokotów,



Rysunek 6.10: Dane pogodowe dla Warszawy

- Ochota,
- Praga Południe,
- Praga Północ,
- Śródmieście,
- Wola,
- Żoliborz.

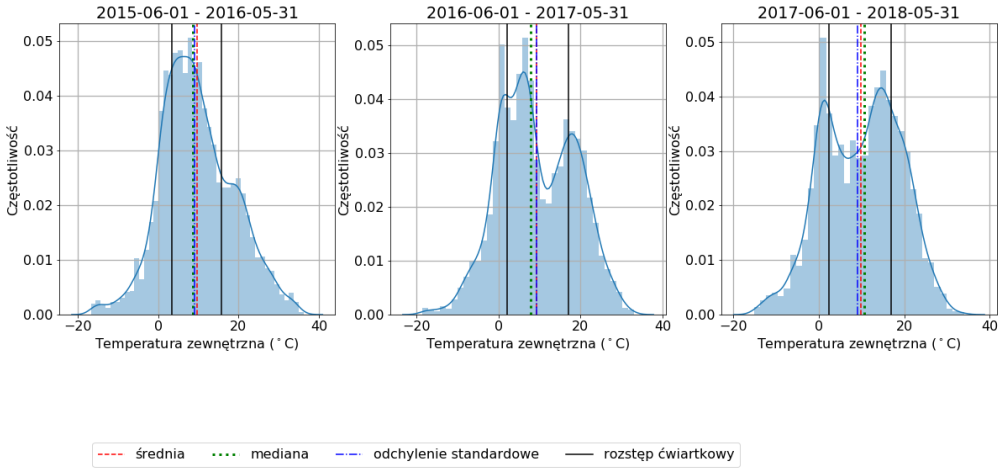
Analiza danych pogodowych w pierwszym etapie sprawdza kompletność tych danych, w przypadku wystąpienia braków w danych pogodowych wartości te nie są uzupełniane, ponieważ stanowią bezpośrednie wartości, na podstawie których trenowany jest model prognostyczny. Walidacja danych pogodowych w pierwszym kroku wymaga wizualnego sprawdzenia danych, następnie analitycznie wyszukiwane są wartości odstające, np. zbyt wysokie lub niskie wartości. Dane pogodowe wykorzystywane w niniejszej pracy otrzymywane są z serwisów informacyjnych, gdzie etap walidacji danych jest uwzględniony, dlatego nie jest wymagana zaawansowana walidacja tychże danych.

Wykresy wykorzystywanych parametrów pogodowych dla Warszawy przedstawione zostały na wykresie 6.10.

Kolejnym etapem analizy danych pogodowych jest analiza rozkładów parametrów w kontekście powtarzalności. By model mógł prawidłowo prognozować zapotrzebowanie na ciepło, ważne jest, aby dane pogodowe wykorzystane w procesie trenowania modelu pokrywały możliwie wszystkie warunki pogodowe, a w szczególności warunki skrajne, takie jak upały czy silne mrozy. W przypadku wykorzystania zaawansowanych modeli uczenia maszynowego, takich jak sieci neuronowe, skuteczność oraz pewność prognozowania maleje wraz z wyjściem poza obszar danych treningowych. Porównanie rozkładów temperatury zewnętrznej dla Warszawy dla trzech analizowanych lat przedstawione zostały na wykresie 6.11. Zauważalny jest podobny zakres temperatur dla analizowanych lat, zmienny jest natomiast kształt rozkładu. Pierwsze dwa lata (lipiec 2015 –lipiec 2017) pod względem kształtu krzywej gęstości rozkładu różnią się od trzeciego roku brakiem wyraźnego rozdziału (dołka) między wyższymi, letnimi temperaturami oraz niższymi, zimowymi temperaturami, oraz miejscem wystąpienia dołka rozkładu na lewo od wartości średniej w przypadku okresu lipiec 2016 –lipiec 2017, oraz na prawo od wartości średniej w przypadku okresu lipiec 2017lipiec 2018. Zróżnicowanie danych ocenione jest jako wystarczające do budowania modelu.

6.2.4 Podział na sezony

Prezentowany w niniejszej pracy system prognostyczny generuje prognozy zapotrzebowania na ciepło przez cały rok. W większości przypadków pracy dużych sieci ciepłowniczych występują trzy typowe okresy, nazywane dalej sezonami, w ciągu roku wpływające na zarządzanie siecią ciepłowniczą, które są przedstawione w definicjach 4,5,6.



Rysunek 6.11: Rozkład temperatury zewnętrznej dla Warszawy.

Definicja 4. Sezon letni jest to okres w którym zapotrzebowanie na ciepło na potrzeby centralnego ogrzewania (c.o.) nie występuje, tzn. wymienniki ciepła na potrzeby c.o. są stale odłączone.

Definicja 5. Sezon zimowy, nazywany również sezonem grzewczym jest to okres w którym wymienniki ciepła na potrzeby c.o. są stale uruchomione.

Definicja 6. Sezon przejściowy jest to okres, w którym wymienniki ciepła na potrzeby c.o. są uruchamiane w odcinkach czasu. O uruchomieniu wymiennika ciepła decyduje system automatyki danego węzła cieplnego na podstawie lokalnej temperatury zewnętrznej oraz administrator budynku.

Sezon przejściowy występuje zawsze między sezonem letnim i zimowym, i charakteryzuje się dużą dynamiką zapotrzebowania na ciepło. Wynika to z nieregularnej pracy wymienników c.o., w momencie gdy pobierane jest ciepło również na cele grzewcze. Wtedy poziom zużycia ciepła jest znacznie wyższy niż w przypadku gdy dla tego samego okresu pobierane jest ciepło wyłącznie na cele ciepłej wody użytkowej.

Podział horyzontu danych na sezony dokonywany jest z wykorzystaniem specjalnego algorytmu, który wymaga zdefiniowania temperatury automatycznego włączenia wymienników c.o.. Temperatura ta często nie jest jednakowa dla wszystkich wymienników w sieci, a odnosi się do znacznej części tych wymienników. Zważywszy na umowny podział horyzontu na sezony możliwy jest dobór temperatury granicznej podziału na sezony jako minimalna temperatura

Tabela 6.5: Warunki podziału danych na sezony.

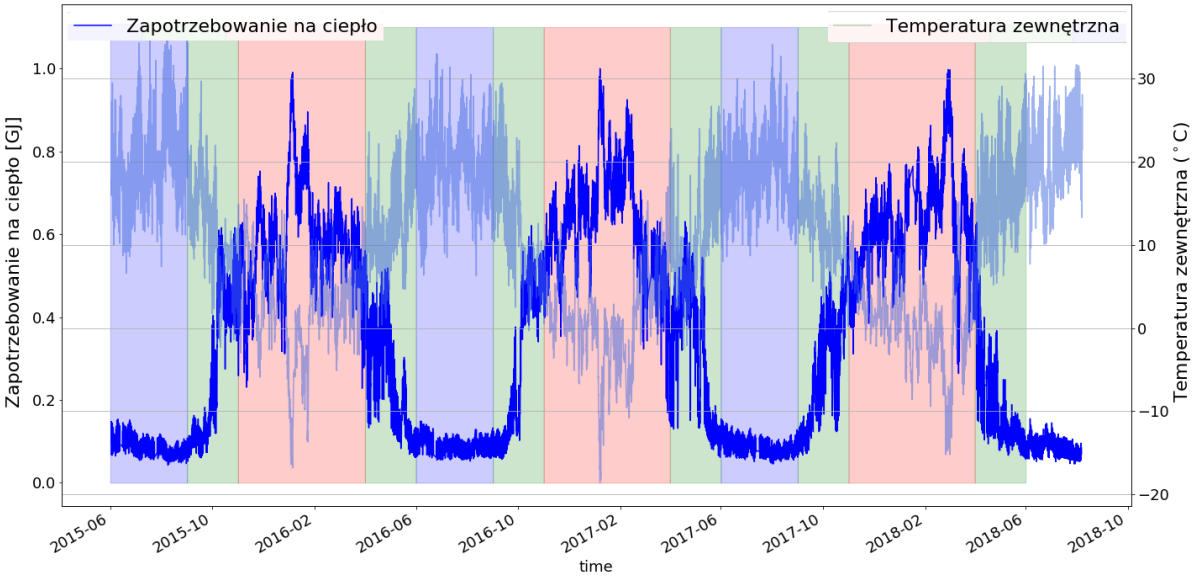
Miesiąc	Warunek	SEZON
Grudzień, Styczeń, Luty	(bezwarunkowo)	ZIMA
Kwiecień, Październik	(bezwarunkowo)	PRZEJŚCIOWY
Czerwiec, Lipiec, Sierpień	(bezwarunkowo)	LATO
Marzec, Listopad	Temperatura minimalna przez kolejne 4 doby <=12 °C	ZIMA
	Pozostałe przypadki	PRZEJŚCIOWY
Od 1 do 15 maja, Od 15 do 30 września	Temperatura minimalna przez 4 kolejne doby > 12 °C	LATO
	Pozostałe przypadki	PRZEJŚCIOWY
Od 16 do 31 maja, Od 1 do 14 września	Średniodobowa temperatura przez kolejne 4 doby > 14 °C	LATO
	Pozostałe przypadki	PRZEJŚCIOWY

uruchomienia wymienników na potrzeby c.o. lub ekspercki dobór tegoż parametru.

W przypadku analizowanej sieci ciepłowniczej ustalono temperaturę graniczną $T_{c.o.}$ jako $12^{\circ}C$, która jest temperaturą graniczną dla większości wymienników c.o. przyłączonych do sieci. Warunki podziału na sezony przedstawiono w tabeli 6.5, natomiast wykres prezentujący zapotrzebowanie na ciepło z zaznaczonymi obszarami poszczególnych sezonów oraz temperaturą zewnętrzną przedstawiony jest na rysunku 6.12. Przedstawione dane z uwagi na anonimizację danych wrażliwych przeskalowane zostały do zakresu [0,1] wyznaczone parametry skalowania (parametry skalowania WSC) wykorzystane zostały w dalszej części pracy do stworzenia wykresów przedstawiających historyczne lub prognozowane zapotrzebowanie na ciepło dla warszawskiej sieci ciepłowniczej. Kolorem niebieskim oznaczony jest obszar sezonu letniego, kolorem czerwonym obszar sezonu zimowego, natomiast kolorem jasnoczerwonym obszar sezonu przejściowego.

6.2.5 Lista bazowa użytkowników sieci

Głównym elementem systemu jest prognoza zapotrzebowania na ciepło dla całej sieci rozumiana jako suma pomiarów zużycia ciepła przez wszystkich odbiorców sieci opisana w rozdziale 3.1. Cechą charakterystyczną dużych sieci ciepłowniczych, takich jak WSC, jest zmienna w czasie liczba odbiorców wynikająca w głównej mierze ze stopniowego przyłączania



Rysunek 6.12: Zapotrzebowania na ciepło dla całej sieci z podziałem na sezony.

nowych odbiorców do sieci, np. przyłączania nowo powstałych osiedli.

Z uwagi na zmienną w czasie liczbę odbiorców sieci stworzona została koncepcja listy bazowej (definicja 7) która pozwala na operowanie stałą liczbą odbiorców.

Definicja 7. Lista bazowa jest to lista użytkowników sieci ciepłowniczej pokrywająca minimum 80 proc. wszystkich odbiorców sieci. Ponadto odbiorcy sieci przyporządkowani do listy bazowej są to odbiorcy w sieci, dla których dostępna jest wymagana historia danych pomiarowych. Długość wymaganej historii danych pomiarowych jest zależna od ogólnej dostępności danych pomiarowych, jednocześnie nie może to być mniej niż jeden pełny rok danych pomiarowych.

W świetle wprowadzonej definicji listy bazowej pojęcie zapotrzebowania na ciepło dla całej sieci rozumie się jako:

Definicja 8. Zapotrzebowanie na ciepło dla całej sieci jest to suma pomiarów zużycia ciepła odbiorców z listy bazowej, w danym punkcie czasu t, wyznaczana według formuły 6.1.

$$ZC(t) = \sum_{i=1}^{n} Q_i(t) \tag{6.1}$$

gdzie:

- n - ilosć odbiorców znajdujących się na liście bazowej,
- Q_i - konsumpcja ciepła przez odbiorcę i.

6.2.6 Wyznaczanie listy bazowej

Zapotrzebowanie na ciepło dla całej sieci zgodnie z definicją 7 odnosi się do odbiorców przyporządkowanych do listy bazowej, a przyporządkowanie to może się odbywać dwojako:

1. Eksperckie wyznaczenie odbiorców sieci, którzy spełniają warunek przyporządkowania do listy bazwej,
2. Wyznaczenie odbiorców do listy bazowej na podstawie dostępnej historii danych pomiarowych.

Prezentowany w niniejszej pracy system zawiera algorytm automatycznie przyporządkowujący odbiorców do listy bazowej. Kryterium przyporządkowania do listy bazowej odnosi się do zakresu danych historycznych, tj. odbiorców, dla których dostępna jest historia danych pomiarowych w okresie nie krótszym niż sześć miesięcy. Dla WSC stanowi to około 83 proc. wszystkich odbiorców sieci zarejestrowanych w systemie bilingowym.

6.2.7 Agregacja danych oraz wyznaczanie zapotrzebowania na ciepło dla całej sieci

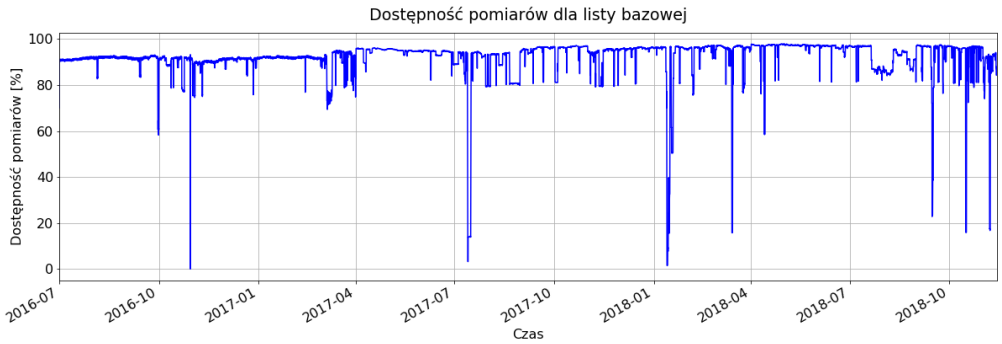
Analizowane dane zawierające pomiary z wymienników ciepła zawierają brakujące rekordy, które w przypadku analizy szeregów czasowych wymagają uzupełnienia, aby zachować ciągłość danych w czasie. Możliwe źródła braków w danych to:

- zmienna liczba odbiorców w czasie, która wynika z faktu, że nie każdy węzeł cieplny posiada historyczne dane pomiarowe dla pełnego horyzontu danych treningowych;
- niedostępności wszystkich wymaganych pomiarów w momencie wyznaczania prognozy, braki mogą powstawać w wyniku awarii lub serwisu czujnika;
- proces walidacji danych, podczas którego część pomiarów jest usunięta.

W przypadku modelowania szeregów czasowych z wykorzystaniem modeli autoregresyjnych brakujące rekordy dzielą zestaw danych na mniejsze zestawy, które należy analizować osobno w wyniku niemożności wyznaczenia cech opóźnionych. W związku z powyższym, jeżeli jest to możliwe, ważne jest przeprowadzenie procesu uzupełniania danych. Przykład macierzy danych po procesie walidacji dla danych pomiarowych zużycia ciepła przedstawiony jest w równaniu 6.2.

$$\begin{bmatrix} 2016-08-2801:00:00 & 0,02990 & 0,151 & 0,0200 & 0,0100 & \dots \\ 2016-08-2802:00:00 & 0,00999 & 0,0091 & 0,0199 & 0,0100 & \dots \\ 2016-08-2803:00:00 & 0,0099 & 0,0090 & 0,200 & 0,0099 & \dots \\ 2016-08-2804:00:00 & NaN & NaN & 0,0099 & 0,0100 & \dots \\ 2016-08-2805:00:00 & 0,0099 & NaN & NaN & 0,0100 & \dots \\ 2016-08-2806:00:00 & 0,0099 & NaN & NaN & NaN & \dots \\ 2016-08-2807:00:00 & 0,0099 & NaN & 0,099 & 0,0100 & \dots \\ \dots & \dots & \dots & \dots & \dots & \end{bmatrix} \tag{6.2}$$

Wyznaczając zapotrzebowanie na ciepło dla całej sieci według równania 6.1 bez uzupełniania brakujących danych dla każdego punktu w czasie zastosowana jest zmienna liczba pomiarów w zależności od ich dostępności. Procent dostępnych pomiarów w funkcji czasu dla analizowanej sieci przedstawiony jest na wykresie 6.13, zauważalne jest, że poziom brakujących pomiarów jest zmienny w czasie i niemożliwe jest zastosowanie stałego współczynnika korekty, dodatkowo dla wybranych punktów w czasie dostępność pomiarów jest nieakceptowalnie niska oraz występują krótkie okresy zaniżonej dostępności na poziomie 80 proc.



Rysunek 6.13: Procent dostępnych pomiarów zużycia energii cieplnej w funkcji czasu.

Koniecznym jest wykonanie uzupełnienia lub przeskalowania brakujących pomiarów, gdyż w przypadku pominięcia kroku uzupełniania brakujących danych wyznaczone zapotrzebowanie będzie pomniejszone o brakujące pomiary, dając nierzeczywistą wartość poziomu zapotrzebowania na ciepło w sieci. Możliwe są dwie metody uzupełniania brakujących danych:

1. Uzupełnienie brakujących pomiarów indywidualnie dla każdego pomiaru/węzła cieplnego. Przy takim podejściu możliwe metody to przykładowo interpolacja brakujących danych z wykorzystaniem najbliższych dostępnych pomiarów, wyznaczenie prognozy zapotrzebowania na ciepło dla danego węzła, a następnie uzupełnienie brakujących danych wynikami prognozy, uzupełnienie brakujących pomiarów stałą wartością, np. średnia wartość zużycia ciepła dla danego sezonu.

2. Przeskalowanie brakujących pomiarów po przeprowadzonym działaniu sumowania zużycia ciepła. W takim przypadku należy stworzyć/wyznaczyć współczynnik przeskalowania, który będzie odpowiadać szacowanemu procentowi zapotrzebowania na ciepło, które nie jest uwzględnione w zestawie danych.

Z przedstawionych powyżej metod uzupełniania brakujących danych, najprecyzyjniejszą wydaje się być metoda wykorzystująca prognozy zużycia ciepła indywidualnie dla każdego odbiorcy. Jest to metoda najmniej efektywna z punktu widzenia funkcjonalności systemu w warunkach rzeczywistych, która wymaga uprzedniego stworzenia modeli prognostycznych dla wszystkich indywidualnych odbiorców. Czas stworzenia modeli oraz uzupełninia brakujących pomiarów dla macierzy danych zawierającej przykładowo dwa tygodnie danych pomiarowych z rozdzielczością godzinową dla kilkunastu tysięcy wymienników ciepła w zależności od dostępnych zasobów sprzętowych może trwać nawet kilkaset godzin. Jest to nieakceptowalny czas z punktu widzenia funkcjonowania całego systemu SWD, którego prezentowany system prognostyczny jest integralną częścią. Konieczne w takim przypadku jest wypracowanie rozwiązania pośredniego, które będzie wystarczająco wydajne czasowo i precyzyjne.

6.2.8 Dynamiczne skalowanie brakujących pomiarów

W celu poprawy efektywności procesu skalowania brakujących pomiarów zaproponowano stworzenie współczynnika pozwalającego na przeprowadzenie dynamicznego skalowania zapotrzebowania na ciepło dla całej sieci oraz obszarów sieci. Współczynnik skalowania dynamicznego *dsc* wyznaczany jest dla każdego punkty w czasie zgodnie z równianiem 6.3. Całkowite zapotrzebowanie na ciepło jest zatem obliczane wedle równania 6.4.

$$dsc(t) = \frac{nb_{total}(t)}{nb_{total\_avaliable}(t)} \tag{6.3}$$

,gdzie

- $nb_{total}(t)$ to liczba węzłów bazowych,
- $nb_{total\_avaliable}(t)$ to całkowita liczba węzłów znajdujących się na liście bazowej, dla której dostępne są pomiary.

$$THD(t) = dsc(t) * \sum_{i=1}^{n=base} h_i(t) \tag{6.4}$$

gdzie

- *base* to liczba węzłów bazowych,
- $h_i(t)$ jest zmierzonym zużyciem ciepła, gdy brakujące wartości uzupełniono wartością zero.

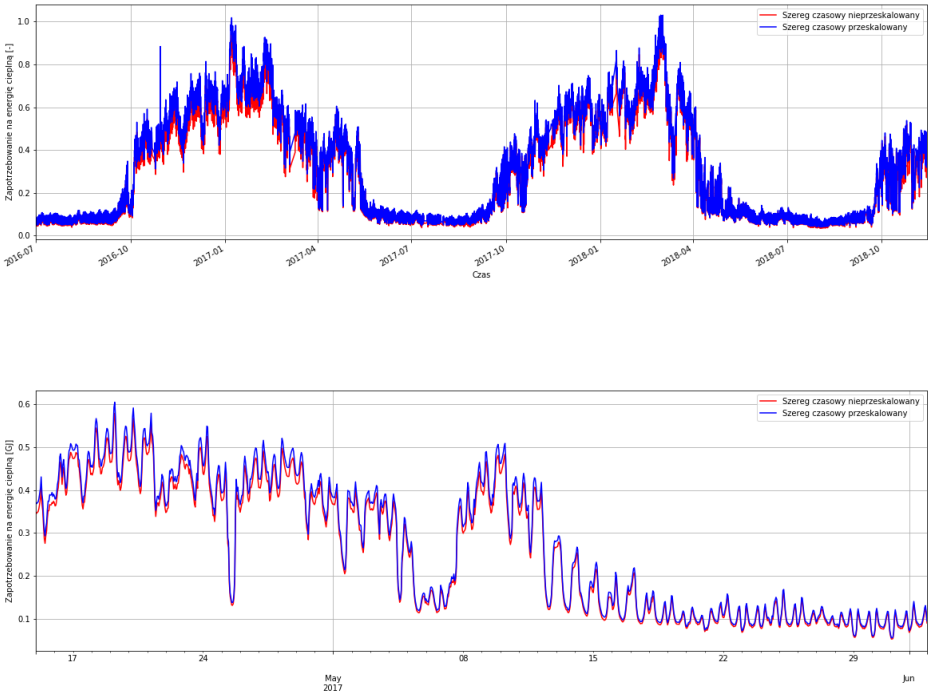
Na wykresie 6.14 przedstawione jest zapotrzebowanie na ciepło w sieci przeskalowane oraz nieprzeskalowane. Przybliżony fragment okresu przejściowego oraz letniego pokazuje skuteczność działania algorytmu również w przypadku sezonu przejściowego, gdzie skoki poziomu zapotrzebowania na ciepło, w szczególności krótkie, kilkugodzinne niskie wartości, mogą być postrzegane jako brakujące dane.

Przeprowadzono dodatkową analizę w celu wyznaczenia błędu skalowania. Z kompletnej macierzy danych losowo wybrane rekordy przypisano jako brakujące wartości, tworząc w ten sposób uszkodzoną macierz, a następnie oszacowano błąd skalowania według wzoru 6.5.

$$bd = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{THD_{NonMV} - \hat{THD}_{MV}}{THD_{NonMV}} \right| \tag{6.5}$$

,gdzie

- THD_{MV} to całkowite zapotrzebowanie na ciepło obliczone na podstawie uszkodzonej macierzy obejmującej dynamiczne skalowanie,



Rysunek 6.14: Porównanie zapotrzebowania na ciepło w sieci przeskalowane oraz nieprzeskalowane.

- THD_{NonV} to całkowite zapotrzebowanie na ciepło obliczone z pełnej, nieuszkodzonej macierzy.

Na postawie powyższej analizy oszacowano błąd skalowania w przedziale 0,8 – 1 proc. dla agregacji zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej. Finalnie, rekordy ze współczynnikiem skalowania wyższym niż 1,25 (sytuacja gdy brakuje ponad 20 % pomiarów) są wykluczane z zestawu danych.

Idea dynamicznego skalowania zapotrzebowania na ciepło może zostać również wykorzystana w przypadku awarii sieci wymagającej wyłączenia części odbiorców. W takim przypadku możliwe jest przeskalowanie całkowitego zapotrzebowania na ciepło do listy bazowej pomniejszonej o liczbę wyłączonych węzłów.

Ponadto przedstawiona powyżej idea dynamicznego skalowania brakujących danych pozwala na estymację zapotrzebowania na ciepło dla wszystkich odbiorców w sieci, wówczas wykorzystywana jest informacja, np. z systemów bilingowych przedsiębiorstwa zawierająca dane o aktualnej, całkowitej liczbie aktywnych węzłów cieplnych w sieci. Zaletą przedstawionej metody jest możliwość uwzględnienia odbiorców, którzy nie zostali jeszcze wyposażeni w system telemetrii pod warunkiem, że liczba tychże odbiorców stanowi mały

odsetek wszystkich użytkowników sieci. Należy mieć na uwadze, że wraz z rozwojem, a tym samym przyłączaniem nowych odbiorców do danej sieci ciepłowniczej, współczynnik skalowania będzie systematycznie wzrastał i w pewnym momencie może przekroczyć poziom akceptowalnie niskiego błędu skalowania. W takim przypadku następuje konieczność aktualizacji listy bazowej, co wiąże się z rekalibracją modelu prognostycznego na podstawie danych pomiarowych uzupełnionych o dane dla nowo uwzględnionych węzłów cieplnych.

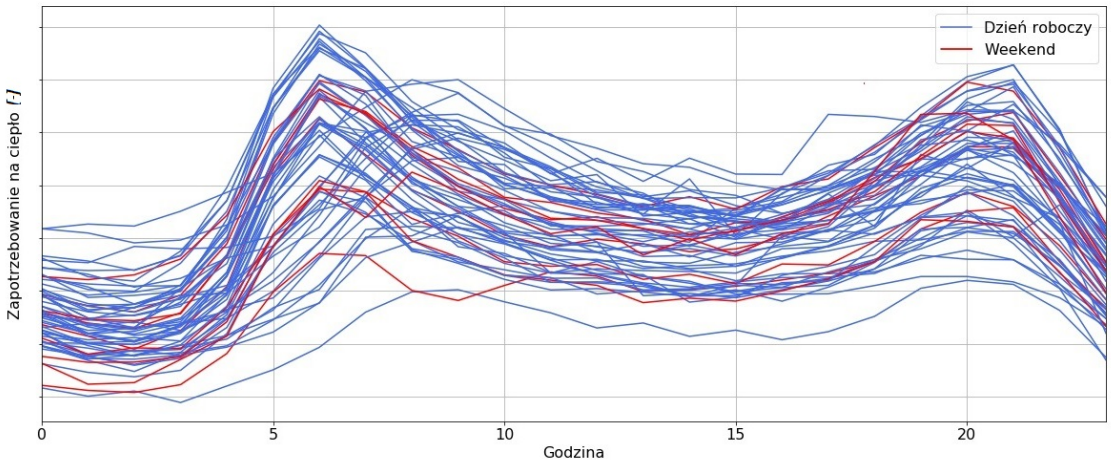
6.2.9 Inżynieria cech

Dane pomiarowe przedstawione w sekcji 6.2.1 stanowią cechy wejściowe do modeli prognostycznych, jednakże jak przedstawiono w sekcji 4.2.2 przydatne jest stworzenie dodatkowych cech, które mogą wpływać na poprawę modelu prognostycznego, stanowiąc dodatkową, cenną informację dla modelu. W kontekście zapotrzebowania na ciepło takimi dodatkowymi cechami, które potencjalnie mogą wyjaśnić zachowanie odbiorców odzwierciedlone w zmiennej zależnej, tj. zapotrzebowaniu na ciepło są:

- zmienne kalendaryczne - parametry zawierające informację o aktualnej chwili w czasie, np. dana godzina, dzień tygodnia, typ dnia itd.;
- transformacje cech pogodowych - stanowią kombinację cech tworzonych, aby poprawić jakość modelowania, np. nieliniowe transformacje cech oraz ekspercko stworzone cechy, które według badań lub wiedzy uzyskanej w wyniku analizy pracy sieci czy analizy dostępnych badań mogą stanowić dobrą reprezentację zmiennej zależnej;
- dane przesunięte w czasie – w przypadku szeregów czasowych, możliwa jest implementacja modelu autoregresyjnego, który wyznacza aktualne wartości prognozowane również w oparciu o wartości historyczne zmiennych zarówno zależnych, jak i niezależnych. W tym celu tworzone są cechy, które są przesunięte w czasie względem cech pierwotnych (danych surowych oraz ich transformacji).
  Tworzone cechy oraz metodologia ich wyznaczania przedstawiona została w dalszej części rozdziału.

Dane kalendaryczne

Główna składowa zapotrzebowania na ciepło, czyli zapotrzebowanie na ciepłą wodę użytkową jest mocno zależna od pory dnia. Analizując zapotrzebowanie na ciepło w sezonie letnim zauważalny jest powtarzalny dzienny schemat konsumpcji ciepła, w którym występuje poranny i wieczorny szczyt zapotrzebowania na ciepło. Wykres 6.15 przedstawia nałożone na siebie dni dla dwóch letnich miesięcy, lipiec oraz sierpień 2015 roku, gdzie zauważalny jest szczyt poranny w okolicach godziny 8.00 oraz szczyt wieczorny w okolicach godziny 21.00.

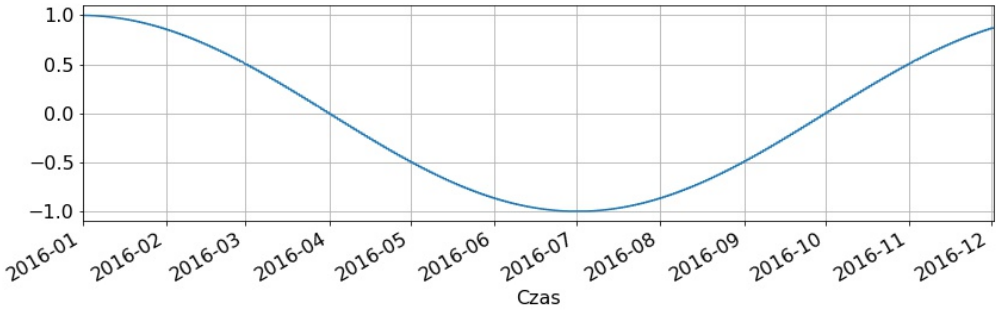


Rysunek 6.15: Znormalizowany wykres dobowego zapotrzebowania na ciepło dla okresu lipiec-sierpień 2015.

Z uwagi na powyższą zależność zapotrzebowania na ciepło od czasu, tworzone są cechy kalendaryczne, które reprezentują w cechach wejściowych czas. Utworzenie wspomnianych cech pozwala na skuteczniejsze modelowanie zachowań, które są powiązane z porą dnia, tygodnia czy roku. Stworzone cechy kalendaryczne przedstawione zostały w tabeli 6.6. Cechy kalendaryczne: godzina, dzień, miesiąc, pora dnia powstają z wykorzystaniem metody kodowania 1 z n (ang. one hot encoding), traktując wskazane elementy daty jako kategorie. Cecha *cos_dzien* została wyznaczona zgodnie z formułą 6.6, gdzie zmienna i odpowiada numerowi porządkowemu dnia roku. Wykres zmiennej dla pełnego roku przedstawiony jest na wykresie 6.16

$$cos\_yday(i) = cos(\frac{i * 2 * \pi}{366}) \tag{6.6}$$

,gdzie i = 1,2,...,366



Rysunek 6.16: Wykres wartości zmiennej *cos_ydzien* na przestrzeni roku.

Tabela 6.6: Tworzone cechy kalendaryczne

Kategoria	**Stworzone cechy**
Godzina	godzina1, godzina2, godzina3, godzina4, godzina5, godzina6, godzina7, godzina8, godzina9, godzina10, godzina11, godzina12, godzina13, godzina14, godzina15, godzina16, godzina17, godzina18, godzina19, godzina20, godzina21, godzina22, godzina23
Dzień	dzien1, dzien2, dzien3, dzien4, dzien5, dzien6
Miesiąc	miesiac1, miesiac2, miesiac3, miesiac4, miesiac5, miesiac6, miesiac7, miesiac8, miesiac9, miesiac10, miesiac11
Typ dnia	dzien_roboczy, weekend, swieto
Dodatkowe	*cos_ydzien*

Przykładowo dla cechy kodującej dzień tygodnia jako kategorię traktuje się numer dnia tygodnia ([1,2,3,4,5,6,7]), w takim przypadku wyróżnia się siedem kategorii, które kodowane wybraną metodą tworzą sześć cech o reprezentacji binarnej. Tworzone cechy kalendaryczne przedstawione zostały w tabeli 6.6, natomiast fragment macierzy danych zawierającej wybrane cechy kalendaryczne przedstawiony został w równaniu 6.7. Cecha *swieto* przyjmuje wartość 1 gdy dany krok odpowiada dniom wymienionych poniżej świąt:

- Nowy Rok, Trzech Króli,
- Niedziela oraz Poniedziałek Wielkanocny,
- Boże Ciało,
- Święto Pracy 1 maja oraz Święto Konstytucji Trzeciego Maja,
- Święto Wojska Polskiego – 15 sierpnia,

- Święto Wszystkich Świętych – 1 listopada,
- Święto Odzyskania Niepodległości –11 listopada,
- Wigilia – 24 grudnia oraz święta Bożego Narodzenia 25–26 grudnia,
- Sylwester.

$$\begin{bmatrix} indeks & godzina1 & godzina2 & godzina3 & ... & dzien1 & dzien28 & ... \\ 2016-01-01\ 01:00:00 & 1 & 0 & 0 & ... & 1 & 0 & ... \\ 2016-08-28\ 01:00:00 & 1 & 0 & 1 & ... & 0 & 1 & ... \\ 2016-08-28\ 02:00:00 & 0 & 1 & 0 & ... & 0 & 1 & ... \\ 2016-08-28\ 03:00:00 & 0 & 0 & 1 & ... & 0 & 1 & ... \\ 2016-08-29\ 04:00:00 & 0 & 0 & 0 & ... & 0 & 0 & ... \\ 2016-12-24\ 07:00:00 & 0 & 0 & 0 & ... & 0 & 0 & ... \\ ... & ... & ... & ... & ... & ... & ... & ... \end{bmatrix} \tag{6.7}$$

Transformacje danych pogodowych

Zbiór potencjalnych cech wejściowych modelu rozszerzany jest również o transformację cech pogodowych. Tworzone cechy wspomagają modele liniowe w przypadku występowania nieliniowych zależności między parametrami pogodowymi, a zmienną zależną. Ponadto tworzone są cechy transformujące pomiar kierunku wiatru do postaci możliwej do wykorzystania przez model. Finalnie tworzona jest dodatkowa cecha, temperatura odczuwalna, która wedle badań oraz wiedzy eksperckiej jest jednym z czynników wpływających na zachowanie konsumentów w kontekście zużycia ciepła. Tworzone cechy oraz ich definicje przedstawione zostały w tabeli 6.7.

Finalnie tworzone są również cechy reprezentujące średnią kroczącą cech pogodowych oraz ich transformacji. Średnia krocząca jest obliczana według wzoru:

$$x\_mean3(t) = \frac{x(t)+x(t\text{-}1\ hour)+x(t\text{-}2hours)}{3} \tag{6.8}$$

gdzie: $x(t)$ to zmierzona wartość dla chwili t, $x\_mean3(t)$ to średnia krocząca cechy x

Tabela 6.7: Transformacje danych pogodowych.

cecha	opis
temperatura∧2	temperatura zewnętrzna^2
temperatura∧3	temperatura zewnętrzna^3
temperatura_sqrt	$\begin{cases} \sqrt{\text{temperatura zewnętrzna}} & \text{jeżeli temperatura zewnętrzna} \geq 0 \\ -\sqrt{\lvert \text{temperatura zewnętrzna} \rvert} & \text{jeżeli temperatura zewnętrzna} < 0 \end{cases}$
wiatr_sqrt	$\sqrt{\text{prędkość wiatru}}$
wiatr∧2	prędkość wiatru^2
wiatr_temp	prędkość wiatru·temperatura zewnętrzna
wiatr_kier_POLNOC	udział kierunku wiatru na osi północnej
wiatr_kier_POLUDNIE	udział kierunku wiatru na osi południowej
wiatr_kier_WSCHOD	udział kierunku wiatru na osi wschodniej
wiatr_kier_ZACHOD	udział kierunku wiatru na osi zachodniej
temp_odczuwalna	$13.12 + 0.6125 \cdot \text{temperatura zewnętrzna} - 11.37 \cdot \text{prędkość wiatru}^{0.16}$ $+0.3965 \cdot \text{temperatura zewnętrzna} \cdot \text{prędkość wiatru}^{0.16}$

Dane przesunięte w czasie: pogodowe oraz zapotrzebowania na ciepło

Powszechnie stosowane oraz wykazujace wysoką dokładność predykcji modele autoregresyjne dla szeregów czasowych wymagają stworzenia cech reprezentujących opóźniane w czasie wartoście zmiennej zależnej, tj. zapotrzebowania na ciepło. Cechy te tworzone są według równania:

$$x_i(t) = x(t-i) \tag{6.9}$$

gdzie i to wartość opóźnienia cechy wyrażona w wielokrotności kroków w czasie danych.
W procesie tworzenia cech opóźnianych określono maksymalne opóźnienie cech, tj. $i_{max} = 50$ *godzin*. Przykładowo dla parametru *Zapotrzebowanie na ciepło* fragment macierzy danych przedstawiony został w róznianiu 6.10

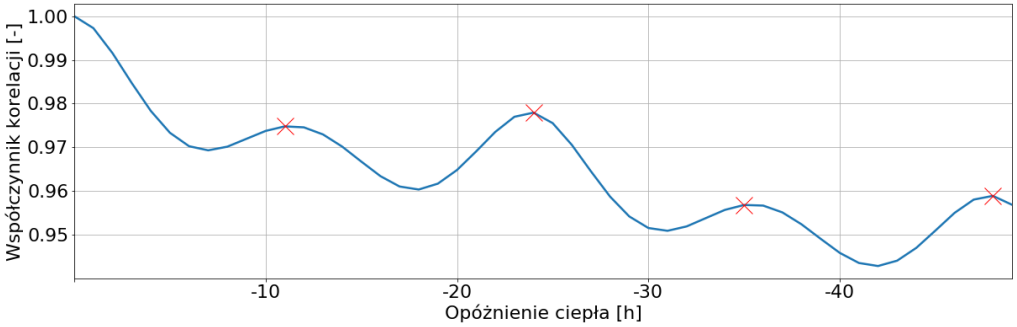
$$\begin{bmatrix} indeks & Cieplo & Cieplo-1 & Cieplo-2 & Cieplo-3 & \ldots \\ 2016-08-2801:00:00 & 7398,29 & NaN & NaN & NaN & \ldots \\ 2016-08-2802:00:00 & 7464,02 & 7398,29 & NaN & NaN & \ldots \\ 2016-08-2803:00:00 & 7593,26 & 7464,02 & 7398,29 & NaN & \ldots \\ 2016-08-2804:00:00 & 7876,65 & 7593,26 & 7464,02 & 7398,29 & \ldots \\ 2016-08-2805:00:00 & 8523,44 & 7876,65 & 7593,26 & 7464,02 & \ldots \\ 2016-08-2806:00:00 & 8569,89 & 8523,44 & 7876,65 & 7593,26 & \ldots \\ 2016-08-2807:00:00 & 7943,67 & 8569,89 & 8523,44 & 7876,65 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \end{bmatrix} \quad (6.10)$$

Dla cech pogodowych oraz ich transformacji również stworzono cechy opóźnione zgodnie z metodologią przedstawioną powyżej.

6.2.10 Analiza autokorelacji zapotrzebowania na ciepło

Autokorelacją nazywamy korelację pomiędzy kolejnymi wartościami tej samej zmiennej. Autokorelacja informuje nas o tym, na ile dane wartości są istotnie związane z obserwacjami zaobserwowanymi wcześniej (o stałym przesunięciu czasowym). Analizowana zmienna, tj. zapotrzebowanie na ciepło dla WSC, przebadana została pod kątem występowania autokorelacji, wykorzystując współczynnik korelacji Pearsona. Wykres 6.17 przedstawia uzyskane wartości współczynnika korelacji zapotrzebowania na ciepło z jej opóźnionymi wartościami. Lokalne maxima uzyskanych wyników oznaczone zostały czerwonymi krzyżykami i wynoszą odpowiednio: 11, 24, 35 i 48 godzin. Wartości opóźniane były co godzinę do maksymalnie 49. godzin. Analiza wykazała istnienie autokorelacji dobowej (lokalne maksimum dla 24 oraz 48 godzin opóźnienia) oraz autokorelację dla połowy doby, gdzie lokalne maximum to 11 godzin opóźnienia.

Przedstawiona analiza pozwala na późniejsze wykorzystanie wyznaczonych opóźnień podczas selekcji cech dla modeli autoregresyjnych.



Rysunek 6.17: Autokorelacja zapotrzebowania na ciepło dla WSC.

6.2.11 Podział danych na zbiór uczący i testowy

Dane przygotowane do trenowania modeli prognostycznych, tj. poddane procesowi walidacji, uzupełniania danych i agregacji oraz zawierające wszystkie dodatkowe cechy/parametry stworzone w ramach procesu inżynieri cech są dzielone na dwa zestawy danch: zbiór uczący/treningowy, oraz zbiór testowy.

Uzyskany szereg czasowy dzielony jest tak,że zbiór testowy zawiera jeden pełny reprezentatywny okres funkcjonowania sieci (dla sieci ciepłowniczych zazwyczaj jest to pełny rok, pokrywając w ten sposób wszystkie sezony: letni, grzewczy, przejściowy między latem a zimą oraz przejściowy miedzy zimą a latem) zawierający najnowsze dane. W przypadku dostępnej długiej historii danych możliwe jest zwielokrotnienie liczby pełnych reprezentatywnych okresów, tak aby zbiór testowy nie przekraczał 30 proc.wszystkich danych.

Zgodnie z powyższym założeniem dla WSC zestaw dostępnych danych został podzielony na zbiór treningowy/uczący oraz testowy zgodnie z następującymi granicznym krokiem czasowym.

- dane dla okresu od 2015-06-01 do 2017-05-31 stanowią zbiór danych treningowych,
- dane dla okresu od 2017-06-01 do 2018-05-01 stanowią zbiór danych testowych.

Powyższy podział dzieli zbiór danych na dwa pełne lata danych treningowych oraz jeden pełny rok danych testowych.

6.2.12 Wyznaczenie najlepszego modelu dla całej sieci

Zakończony proces zbierania oraz przygotowania macierzy danych pozwala na tworzenie i testowanie różnych typów oraz klas modeli prognostycznych. W niniejszym rozdziale

przedstawiona została metodologia tworzenia modeli prognostycznych oraz wyboru modelu przeznaczonego do implementacji w systemie. Model wyselekcjonowany do implementacji w systemie ma za zadanie generację prognoz zapotrzebowania na ciepło zgodnie z konfiguracją w systemie tak, aby możliwe było funkcjonowanie systemów optymalizacji.

Tak jak przedstawiono w rozdziale 3.3, na postawie dostępnych badań naukowych nie jest możliwe jednoznaczne określenie typu oraz klasy najlepszego modelu opartego o uczenie maszynowe prognozującago zapotrzebowanie na ciepło. Prezentowane wyniki badań odnoszą się do ciepła wprowadzanego do sieci przez źródła ciepła lub obejmują prognozy indywidualnych odbiorców. Badania nie obejmują także aspektu implementacji w systemi rzeczywistym, który operuje w trybie online, co nakłada dodatkowe kryteria doboru najlepszego modelu prognostycznego.

Prognoza zapotrzebowania na ciepło jako zmienna wykorzystywana przez optymalizator musi pokrywać horyzont optymalizacji, który wynosi od 24 godzin do 5 dni. Pierwsze 24 godziny są kluczowe w kontekście bieżącej optymalizacji pracy sieci, równie istotne są następujące po nich godziny od 25. do 48. Prognozy w dalszej części horyzontu predykcji od 3. do 5. doby stanowią w znacznej mierze wartość informacyjną dla operatorów sieci. W przypadku wykorzystywania modeli autoregresyjnych należy mieć na uwadze, że dokładność prognozy pogarsza się wraz z długością horyzontu predykcji z uwagi na kumulację błędu prognozy. W związku z powyższym podczas etapu trenowania oraz doboru modelu prognostycznego analizowana była dokładność predykcji z podziałem na sezon oraz zakres horyzontu predykcji, tj. dla każdego sezonu analizowano dokładność predykcji dla:

- pierwszych 24. godzin horyzontu predykcji.
- od 25. do 48. godziny horyzontu predykcji,
- od 49. do 120. godziny horyzontu predykcji.

Z uwagi na powyższe w części analitycznej tworzenia systemu trenowane są różne modele prognostyczne, a następnie w oparciu o zdefiniowane kryteria doboru najlepszego modelu wybierany jest model do implementacji w systemie SWD.

Kryteria doboru najlepszego modelu

1. Dokładność prognozy z rozróżnieniem na sezon letni, przejściowy oraz grzewczy, oraz horyzonty predykcji.

2. Efektywność i czas trenowania modelu.

Poniżej przedstawtione została lista testowanych modeli prognostycznych:

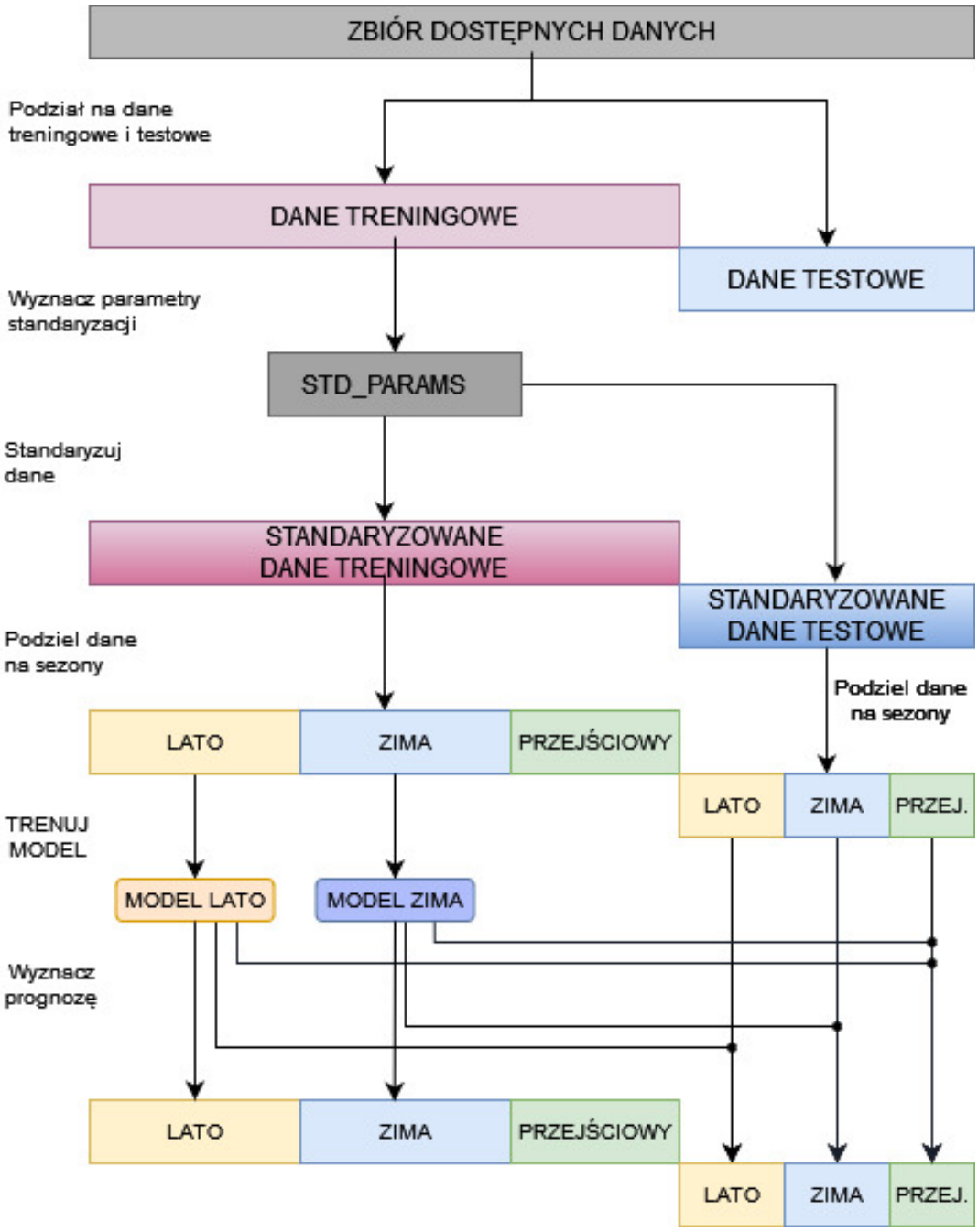
- regresja grzbietowa z zmiennymi niezależnymi (RR) - model z wykorzystaniem autoregresji oraz bez autoregresji;
- model rozmyty dla sezonu przejściowego (FUZZY). Funkcja rozmycia zależna od temperatury zewnętrznej;
- model rozmyty dla sezonu przejściowego (FUZZY). Funkcja rozmycia zależna od temperatury zewnętrznej oraz nasłonecznienia;
- sztuczna sieć neuronowa (SSN);
- sztuczna sieć neuronowa z autoregresyjnym wejściem;
- modele częściowo liniowe, gdzie temperatura zewnętrzna jest podzielona na 3 zmienne niezależne;
- połączenie dwóch modeli: modelu prognozujacego zapotrzebowanie na ciepło na potrzeby centralnego ogrzewania oraz modelu reszt;
- połączenie modeli dla sezonu letniego: modelu dla dni roboczych oraz modelu dla dni weekendowych;
- korekta sezonu przejściowego poprzez dodanie modelu rezyduów;
- Las losowy [12];
- połączenie modeli grup odbiorców o podobnej charakterystyce konsumpcji ciepła klastrowanych wg. metody PDC [23] lub DTW [3].

W rozprawie przedstawione zostaną dwa modele:

- model który uzyskał najlepszą dokładność predykcji, tj. SSN (połącznie modelu z autoregresją oraz bez autoregresji),
- model mniej wymagający obliczeniowo, który uzyskał dobre dokładności predyckji dla sezonu letniego oraz zimowego, tj. regresja grzbietowa (połącznie modelu z autoregresją oraz bez autoregresji).

Regresja grzbietowa (RR)

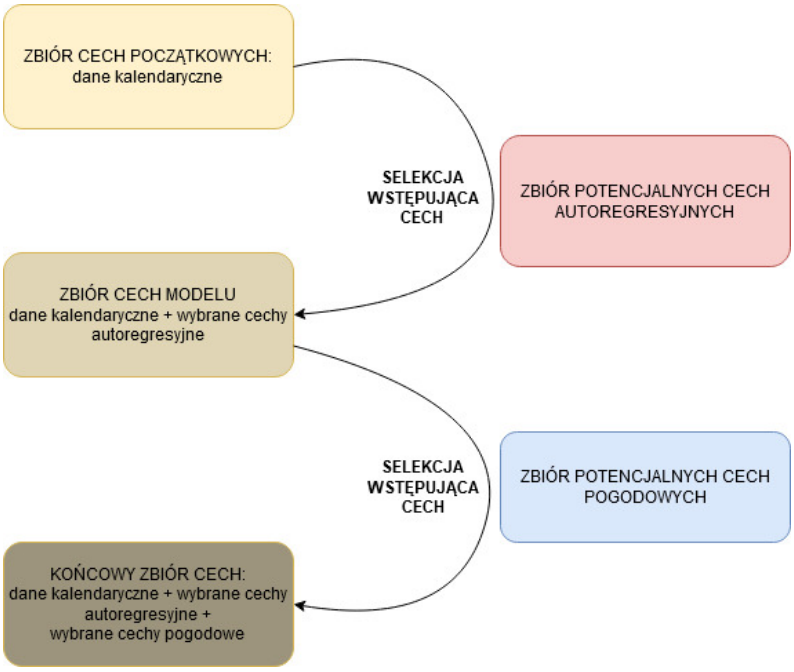
Model RR, którego założenia przedstawioe zostały w sekcji 2.2 trenowany był dla każdego sezonu osobno. Model ten wymaga uprzedniej standaryzacji danych, dlatego na podstawie danych treningowych wyznaczone zostały parametry standaryzacji. Standaryzowane dane



Rysunek 6.18: Metodologia wyznaczania modelu regresji Ridge.

treningowe podzielone na sezony wykorzystane zostały do trenowania modeli indywidualnie dla każdego sezonu. Z uwagi na bardzo dużą liczbę potencjalnych cech pogodowych (192 potencjalne cechy wejściowe) wykonana została selekcja cech wejściowyc modelu

z wykorzystaniem selekcji wstepującej opisanej w sekcji 4.2.2. Hiperparametr modelu, tj. parametr $alfa(\alpha)$ kontorlujący stopień kary wybrany jest metodą siatki, opisanej w rozdziale 4.2.2, spośród wartości $wektor[1e-10, 10]$. Finalna wartosc parametru $alfa$ wynosi $1e-07$ Algorytm podziału danych, trenowania modelu oraz wyznaczania prognozy przedstawiony został na wykresie 6.18. Zbiór dostępnych danych podzielony został na zbiór danych treningowych oraz testowych. Na podstawie zbioru danych treningowych wyznaczone zostały parametry konieczne do przeprowadzenia standaryzacji danych, a następnie na postawie wyznaczonych wartości wykonana została standaryzacja danych treningowych i testowych.



Rysunek 6.19: Metodologia selekcji cech modelu regresji grzbietowej z autoregresją.

Modele dla sezonu letniego oraz zimowego trenowane były z wykorzystaniem standaryzowanych danych treningowych dla danego sezonu, natomiast z uwagi na nieakceptowalną dokładność prognozy modelu trenowanego dla sezonu przejściowego w sezonie tym wykorzystano model rozmyty opisany w dalszej części rozdziału.

Selekcja cech wejściowych modelu z wykorzystaniem selekcji wstępującej dla modelu RR bez autoregresji polegała na dobraniu do cech kalendarycznych odpowiednich cech pogodowych ich transformacji oraz opóźnień. Natomiast w przypadku modelu RR z autoregresyjnym wejściem selekcja cech wykonywana była dwuetapowo. Algorytm selekcji cech zobrazowany został na wykresie 6.19. W pierwszym etapie do cech początkowych metodą selekcji wstępującej dobierane były cech autoregresyjne, tj. opóźnienia zmiennej prognozowanej. Następnie do cech początkowych oraz wybranych cech autoregresyjncyh

metodą selekcji wstępującej dobierane były cechy pogodowe, ich transformacje oraz opóźnienia.

Wybrane cechy wejściowe dla modeli RR przedstawione zostały w tabeli 6.8. Wyniki prognoz dla sezonu letniego oraz zimowego dla modelu z autoregresją przedstawione zostały w tabeli 6.9, natomiast dla modelu bez autoregresji w tabeli 6.12.

Tabela 6.8: Cechy wejściowe modeli RR

	SEZON LETNI		**SEZON ZIMOWY**	
	model RR bez autoregresji	**model RR z autoregresją**	**model RR bez autoregresji**	**model RR z autoregresją**
Cechy kalendaryczne	dzień 1, ..., dzień 6, godzina1, ..., godzina23, miesiąc1, ..., miesiąc 11, weekend,święto	dzień 1, ..., dzień 6, godzina1, ..., godzina23, miesiąc1, ..., miesiąc 11, weekend,święto	dzień 1, ..., dzień 6, godzina1, ..., godzina23, miesiąc1, ..., miesiąc 11, weekend,święto	dzień 1, ..., dzień 6, godzina1, ..., godzina23, miesiąc1, ..., miesiąc 11, weekend,święto
Cechy pogodowe	temperatura, temp_odczuwalna, temp^3-1, temp^3-2, temp_odczuwalna-1, temp_odczuwalna-2, wilgotnosc	temperatura, temp_odczuwalna	temperatura-1, temperatura-48, naslonecznienie-1, temperatura_sqrt, zachmurzenie-10, wilgotnosc-1	temperatura, temperatura-1, wiatr_sqrt, naslonecznienie, naslonecznienie-2, temperatura_sqrt, naslonecznienie-4, kierunek_wiatru-5, wilgotnosc-3
Cechy autoregresyjne	-	cieplo-1, cieplo-23, cieplo-26	-	cieplo-1, cieplo-2, cieplo-21

Model rozmyty dla sezonu przejściowego (FUZZY)

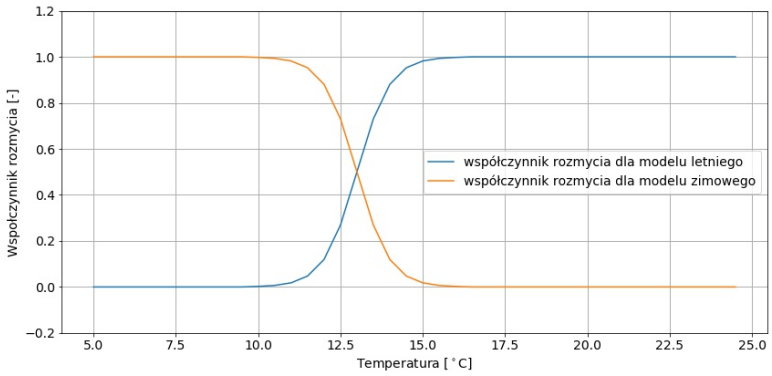
Z uwagi na bardzo słabą dokładność modeli RR w obszarze sezonu przejściowego, prognozę zapotrzebowania na ciepło w tymże okresie wyznaczono przez rozmywanie wyników prognozy z wykorzystaniem modelu dla sezonu letniego oraz zimowego. Model ten wymaga stworzenia optymalnej krzywej lub algorytmu funkcji rozmycia. Podstawową funkcją rozmycia jest funkcja zależna od temperatury zewnętrznej, dla której należy wyznaczyć optymalne wartości graniczne T_1 oraz T_2, pomiędzy którymi następuje rozmycie wyników prognoz. Zakres temperatur, dla których wyznaczane są optymalne wartości zależy w dużej mierze od ustawień automatyki wymienników ciepła. Dla analizowanej sieci temperatura progowa jest ustalona w systemie automatyki i waha się od 10 $^{\circ}C$ do 16 $^{\circ}C$, dlatego wartości przejścia dobierane zostały z zakresu

$T_1 \in < 8,14 >$ oraz $T_2 \in < 14,22 >$. Wartości optymalne wynoszą $T_1 = 10, T_2 = 16$, natomiast funkcję rozmycia przedstawiono na rysunku 6.20.

Ostatecznie wartość prognozy dla modelu rozmytego jest wyznaczana zgodnie z równaniem 6.11.

$$\begin{aligned} Prognoza(t) = coeff\_s(temp(t)) * y_{arx\_lato} \\ + coeff\_h(temp(t)) * y_{arx\_zima} \end{aligned} \tag{6.11}$$

gdzie: coeff_s - współczynnik rozmycia dla modelu letniego, y_{arx_lato} - wynik prognozy z wykorzystaniem modelu letniego, coeff_h - współczynnik rozmycia dla modelu zimowego, y_{arx_zima} - wynik prognozy z wykorzystaniem modelu zimowego.



Rysunek 6.20: Funkcja rozmycia dla modelu rozmytego.

Y_{arx_lato} to model ARX wyznaczony dla sezonu letniego, oznacza to, że dane z sezonu letniego zostały wykorzystane do dopasowania modelu. Analogicznie y_{arx_zima} model ARX dla sezonu zimowego został dopasowany z wykorzystaniem danych dla sezonu zimowego. Wyniki dokładności prognoz dla sezonu przejściowego z wykorzystaniiem modelu letniego i zimowego z autoregresją przedstawiono w tabeli 6.9, natomiast z wykorzystaniem modeli bez autoregresji w tabeli 6.12.

Tabela 6.9: Dokładność prognoz dla całej sieci z wykorzystaniem modelu RR z autoregresją z podziałem na sezony.

			LATO	ZIMA	PRZEJSCIOWY
horyzont predykcji	**score**	**dane**			
1-24 godziny	**MAPE [%]**	**dane treningowe**	5.29	2.80	8.5
		dane_testowe	5.89	3.26	10.96
	MAE [-]	**dane treningowe**	41.94	140.74	209.75
		dane_testowe	48.91	144.78	192.53
25-48 godzina	**MAPE [%]**	**dane treningowe**	5.86	2.98	14.16
		dane_testowe	6.90	4.46	12.55
	MAE [-]	**dane treningowe**	46.25	148.58	239.47
		dane_testowe	52.67	155.30	215.26
49-120 godzina	**MAPE [%]**	**dane treningowe**	6.45	3.56	16.49
		dane_testowe	7.58	4.54	13.77
	MAE [-]	**dane treningowe**	50.44	155.77	269.87
		dane_testowe	58.16	168.17	231.17

Tabela 6.10: Dokładność prognoz dla całej sieci z wykorzystaniem modelu RR z autoregresją z podziałem na miesiące.

	1-24 godziny				**25-48 godzina**				**49-120 godzina**			
	MAPE [%]		**MAE [-]**		**MAPE [%]**		**MAE [-]**		**MAPE [%]**		**MAE [-]**	
	TR	**TS**	**TR**	**TS**	**TR**	**TS**	**TR**	**TS**	**TR**	**TS**	**TR**	**TS**
Styczeń	2.98	2.29	135.17	121.66	2.11	2.42	143.51	128.82	2.19	2.42	147.88	128.82
Luty	3.25	1.85	125.43	124.09	2.31	1.9	128.64	128.46	2.38	1.9	131.67	128.46
Marzec	4.81	3.46	180.9	181.4	5.17	3.94	188.27	202.27	5.67	3.94	197.07	202.27
Kwiecień	7.86	12.78	233.45	266.29	8.58	15.01	253.87	296.48	9.38	15.01	271.6	296.48
Maj	11.23	8.18	179.8	88.75	12.98	9.28	203.98	99.69	14.69	9.28	220.58	99.69
Czerwiec	6.06	6.59	53.56	54.18	6.69	7.23	58.74	59.5	7.12	7.23	61.84	59.5
Lipiec	4.26	6.1	32.92	47.13	4.52	6.3	34.93	48.39	4.51	6.3	35.02	48.39
Sierpień	5.55	6.59	39.72	45.88	6.36	7.29	45.37	51.05	7.65	7.29	54.37	51.05
Wrzesień	12.85	13.63	246.15	204.88	28.16	16.13	304.6	241.46	34.55	16.13	381.58	241.46
Październik	4.87	9.57	171.43	250.7	5.21	10.04	184.15	264.47	5.17	10.04	190.43	264.47
Listopad	3.2	3.4	139.55	153.08	3.46	3.73	149.32	168.69	3.64	3.73	156.51	168.69
Grudzień	2.74	2.73	140.61	145.56	2.93	2.87	150.82	152.22	3.12	2.87	163.89	152.22

Tabela 6.11: Dokładność prognoz dla całej sieci z wykorzystaniem modelu RR bez autoregresji z podziałem na miesiące.

parametr	**MAPE [%]**		**MAE [-]**	
dane	**TR**	**TS**	**TR**	**TS**
Styczeń	2.68	4.13	184.68	215.57
Luty	2.52	1.7	140.41	109.01
Marzec	4.67	3.94	182.31	234.98
Kwiecień	11.08	21.7	306.94	431.61
Maj	15.42	16.35	530.43	196.76
Czerwiec	7.33	7.66	64.12	60.58
Lipiec	4.88	6.92	37.41	52.07
Sierpień	8.22	8.89	57.76	60.44
Wrzesień	14.35	27.37	166.78	579.24
Październik	6.35	8.61	225.88	267.6
Listopad	3.96	3.62	172.38	168.03
Grudzień	3.4	3.25	174.4	174.82

Tabela 6.12: Dokładność prognoz dla całej sieci z wykorzystaniem modelu RR bez autoregresji.

		LATO	**ZIMA**	**PRZEJSCIOWY**
parametr	**dane**			
MAPE [%]	**dane treningowe**	6.74	3.23	14.45
	dane_testowe	7.24	3.43	19.43
MAE [-]	**dane treningowe**	53.35	162.89	320.97
	dane_testowe	55.03	174.24	391.80

Sztuczna sieć neuronowa (SNN)

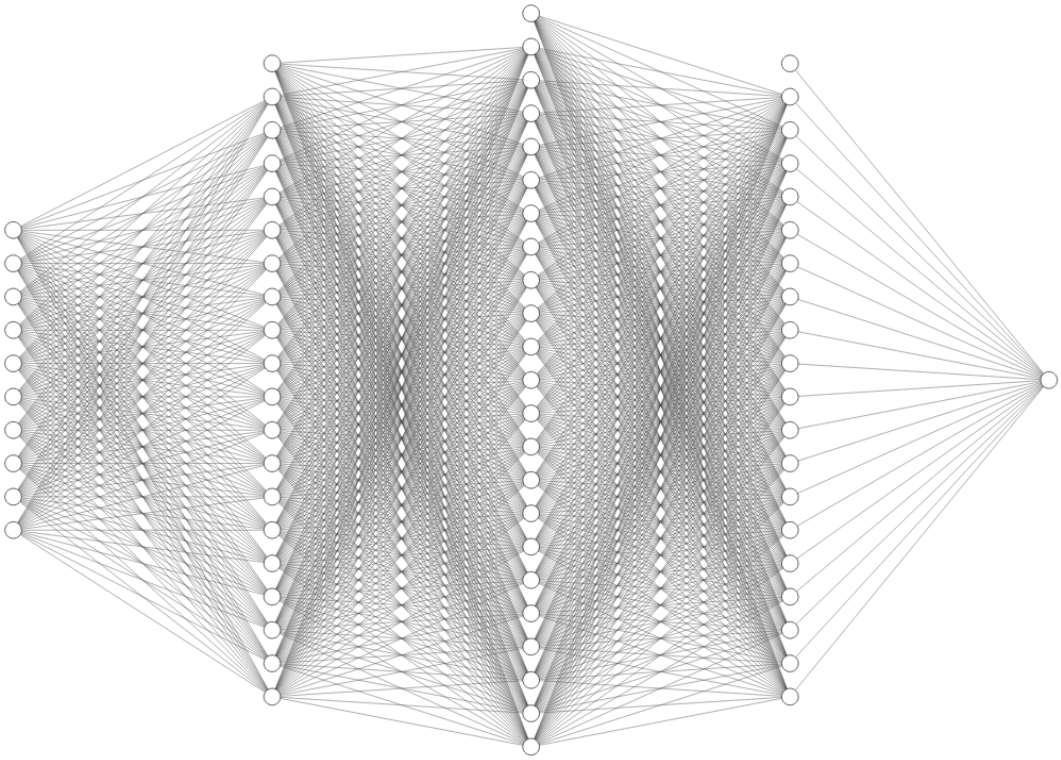
Najlepsze rezultaty uzyskane zostały dla głębokiej sztucznej sieci neuronowej. Zastosowano SSN czterowarstwowe o liczbach neuronów odpowiednio 400x450x400x1. Dla pierwszych trzech warstw zastosowano funkcję aktywacji RELU (Rectified Linear Unit). Dla ostatniej warstwy wykorzystano aktywację funkcją liniową. Wizualizacja wykorzystanej SSN, gdzie warstwy z wyjątkiem warstwy wyjściowej zawierają 20-krotnie mniej neuronów, odpowiednio: 20x23x20x1 przedstawiona została na wykresie 6.21.

Optymalne hiperparametry dla czterowarstwowej SSN wyselekcjonowano przy użyciu metody siatki z walidacją krzyżową. Przetrenowano dwie sztuczne sieci neuronowe:

- SSN z cechami wejściowymi zawierającymi dane pogodowe, dane kalendaryczne oraz ich opóźnione wartości (sztuczna sieć neuronowa skierowana (ang. Feed-Forward) bez wejścia autoregresyjnego);
- SSN z dodatkowymi cechami wejściowymi zawierającymi opóźnione dane zapotrzebowania na ciepło (autoregresyjna SSN z wejściem autoregresyjnym).

W przeciwieństwie do innych analizowanych modeli SSN nie używają definicji sezonu, a przedstawione SSN były trenowane z wykorzystaniem danych dla całego roku, natomiast pozostałe modele tworzone były dla poszczególnych sezonów. Wprowadzona różnica w trenowaniu SSN wynikała z faktu, iż lepsze wyniki uzyskano, jeśli istniał jeden model SSN na cały rok w porównaniu do sytuacji, gdzie istniały modele SSN na każdy sezon.

Wszystkie wagi sieci neuronowych zostały zainicjowane zgodnie z inicjalizacją Xaviera [43]. Architekturę sieci neuronowej uzyskano metodą prób i błędów. Sieć została zoptymalizowana przy użyciu propagacji wstecznej z optymalizatorem szybkości uczenia się Adama i minimalizacji średniego procentowego błędu jako funkcji straty. Model dla określonych cech wejściowych prognozuje jedną cechę wyjściową – prognozę zapotrzebowania na ciepło WSC dla danej godziny [43].



Rysunek 6.21: Wizualizacja zredukowanego modelu SSN dla modelu całej sieci.

Finalnie stworzono trzy modele cząstkowe, które pozwalają na wyznaczenie prognozy zapotrzebowania na ciepło dla WSC w całym wymaganym horyzoncie predykcji dla dowolnego okresu roku:

- model SSN z autoregresyjnym wejściem w pierwszych 24 h (AR-SN);
- model SSN bez autoregresyjnego wejścia wykorzystywany w kolejnych godzinach, 25 h-120 h (SSN);
- model SSN bez autoregresyjnego wejścia wykorzystywany w kolejnych godzinach, 25 h–120 h, wykorzystywany tylko w dniach 15–30.09., (SSN 15-30.09.).

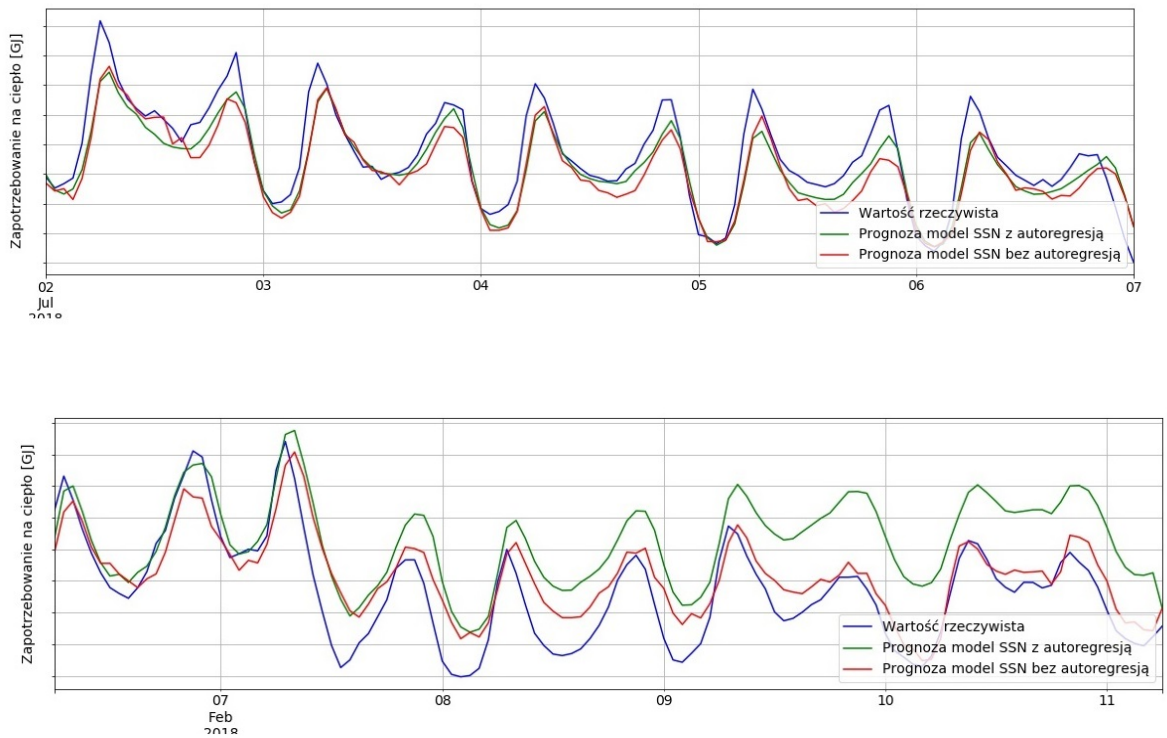
Okres 15–30.09. jest okresem, w którym następuje uruchomienie wszystkich wymienników centralnego ogrzewania po sezonie letnim. Z uwagi na działanie systemów automatyki uruchamiających wymienniki centralnego ogrzewania jest to okres szczególnie wrażliwy na zmiany wartości temperatury zewnętrznej, dlatego koniecznym okazało się stworzenie dedykowanego modelu dla wspomnianego okresu, w przypadku gdy nie wykorzystywano danych autoregresyjnych.

Parametry trenowania poszczególnych modeli przedstawione zostały w tabeli 6.13, natomiast wyniki dla modelu SSN z wykorzystaniem autoregresji w podziale na horyzont

czasowy z podziałem na sezony przedstawione zostały w tabeli 6.16, a z podziałem na miesiące w tabeli 6.15. Wyniki wyznaczane były iteracyjnie dla każdej doby. Godzina startu prognozy była zmienna tak, aby zniwelować wpływ godziny prognozy na wynik predykcji.

Tabela 6.13: Parametry trenowania SSN.

	model AR-SSN	**model SSN**	**model SSN 15-30.09**
liczba epok	7x10	2x8	9
optymalizator	ADAM	ADAM	ADAM
wielkość porcji danych (batch size)	168	168	168
inicjalizacja wag	Xavier	Xavier	Xavier
rodzaj normalizacji danych wejściowych	MinMax	Standaryzacja	Standaryzacja



Rysunek 6.22: Porównanie prognoz dla modelu SSN z autoregresją oraz bez autoregresji.

Wyniki dokładności predykcji dla modelu SSN bez autoregresji z podziałem na sezony przedstawione zostały w tabeli 6.14, natomiast z podziałem na miesiace w tabeli 6.17. Wyniki pokazują, że dla sezonu i miesięcy zimowych dokładność predykcji jest wysoka oraz porównywalna lub lepsza z wynikami przedstawianymi w literaturze opisanymi w rozdziale

3.3 zarówno dla modelu z autoregresją, jak i modelu bez autoregresji. Sezon letni oraz sezon przejściowy również uzyskują dobrą jakość predykcji, natomiast z uwagi na brak dostępnych danych nie jest możliwe porównanie dokładności prognoz modelu z innymi przedstawianymi w literaturze. Wykresy wybranych prognoz zapotrzebowania na ciepło w sezonie zimowym oraz letnim przedstawione zostały na wykresie 6.22. Dla dnia z okresu sezonu zimowego zauważalny jest wyraźny spadek dokładności modelu SSN z autoregresją po pierwszej dobie prognozy w porównaniu do wyników modelu bez autoregresji. Natomiast model SSN bez autoregresji w pierwszej dobie prezentuje mniejszą dokładność prognozy w porównaniu do modelu SSN z autoregresją. Zjawisko to zostało odzwierciedlone w wynikach dokładności prognozy, gdzie dla sezonu zimowego dokładność predykcji z wykorzystaniem modelu SSN z autoregresją dla pierwszej doby wynosi 3,08 proc. oraz 3,91 proc dla dalszej części horyzontu w porównaniu do dokładności predykcji modelu SSN bez autoregresji na poziomie 3,26 proc.. Analogiczne wyniki uzyskano dla sezonu letniego i przejściowego, gdzie dla pierwszej doby dokładność modelu bez autoregresji wynosi 7,58 (sezon letni) oraz 9,58 (sezon przejściowy), natomiast model z autoregresją prezentuje dokładność 6,66 proc. (sezon letni) i 8,41 proc. (sezon przejściowy) dla pierwszej doby oraz 7,87 proc. (sezon letni) i 9,78 proc. (sezon przejściowy) dla doby kolejnej. Dlatego też w przypadku wdrożenia modelu, model w momencie wyznaczania prognozy zapotrzebowania na ciepło przez pierwsze 24 godziny wykorzystywany jest model SSN z autoregresją, natomiast dla kolejnych godzin prognozy wykorzystywany jest model SSN bez autoregresji. Uzyskane wyniki są wystarczające, aby wykorzystać model jako zmienną wyjściową optymalizatora sieci ciepłowniczej. Uzyskane wyniki są wystarczające aby wykorzystać model jako zmienną wyjściową optymalizatora sieci ciepłowniczej.

Tabela 6.14: Dokładność prognoz dla całej sieci z wykorzystaniem modelu SSN bez autoregresji.

		LATO	ZIMA	PRZEJSCIOWY
parametr	**dane**			
MAPE [%]	**dane treningowe**	6.74	2.95	9.03
	dane_testowe	7.58	3.26	9.58
MAE [-]	**dane treningowe**	43.93	152.32	191.01
	dane_testowe	61.69	218.81	211.26

Tabela 6.15: Dokładność prognoz dla całej sieci z wykorzystaniem modelu SSN z autoregresją z podziałem na miesiące. TR - zbiór danych treningowych. TS - zbiór danych testowych.

	1-24 godziny				**25-48 godzina**				**49-120 godzina**			
	MAPE [%]		**MAE [-]**		**MAPE [%]**		**MAE [-]**		**MAPE [%]**		**MAE [-]**	
	TR	**TS**	**TR**	**TS**	**TR**	**TS**	**TR**	**TS**	**TR**	**TS**	**TR**	**TS**
MIESCIĄC												
Styczeń	2.09	2.83	139.83	158.91	2.69	3.3	178.47	185	4.58	3.3	301.49	185
Luty	2.86	2.22	157.08	147.1	3.46	2.66	190.7	177.21	5.19	2.66	289.21	177.21
Marzec	3.35	4.62	143.39	231.87	3.64	5.42	154.09	270.86	4.22	5.42	173.89	270.86
Kwiecień	5.25	10.18	167.39	188.55	5.72	10.94	180.38	200.18	6.27	10.94	192.3	200.18
Maj	6.06	8.36	100.25	88.13	6.66	9.12	108.61	95.48	7.38	9.12	115.2	95.48
Czerwiec	4.94	6.47	41.82	53.01	5.14	6.89	43.26	56.66	5.37	6.89	44.66	56.66
Lipiec	3.6	6.03	27.2	43.83	3.7	6.39	27.9	45.9	3.78	6.39	28.62	45.9
Sierpień	5.28	8.08	36.44	55.47	5.56	8.61	38.31	59.59	5.83	8.61	40.25	59.59
Wrzesień	9.59	8.04	113.35	139.49	11.54	8.92	138.3	158.63	14.75	8.92	184.11	158.63
Październik	3.69	7.08	138.83	196.51	4.07	7.74	154.79	211.74	4.82	7.74	189.25	211.74
Listopad	3.1	3.73	141.85	166.61	3.57	4.6	164.06	205.86	4.74	4.6	221.63	205.86
Grudzień	2.81	2.81	144.46	146.45	3.49	3.57	180.14	185.85	5.12	3.57	270.17	185.85

Tabela 6.16: Dokładność prognoz dla całej sieci z wykorzystaniem modelu SSN z autoregresją z podziałem na sezony.

horyzont predykcji	score	dane	LATO	ZIMA	PRZEJSCIOWY
1-24 godziny	**MAPE [%]**	**dane treningowe**	5.62	2.83	6.11
		dane_testowe	6.66	3.08	8.41
	MAE [-]	**dane treningowe**	35.13	145.21	129.98
		dane_testowe	49.50	169.41	151.92
25-48 godzina	**MAPE [%]**	**dane treningowe**	5.81	3.36	6.95
		dane_testowe	7.87	3.91	9.78
	MAE [-]	**dane treningowe**	36.47	173.81	145.47
		dane_testowe	52.55	204.50	165.20
49-120 godzina	**MAPE [%]**	**dane treningowe**	5.00	4.77	8.24
		dane_testowe	7.94	5.39	10.45
	MAE [-]	**dane treningowe**	37.83	253.04	170.08
		dane_testowe	55.83	282.76	187.17

Tabela 6.17: Dokładność prognoz dla całej sieci z wykorzystaniem modelu SSN bez autoregresja z podziałem na miesiące.

	MAPE [%]		**MAE [-]**	
	TR	**TS**	**TR**	**TS**
Styczeń	2.11	3.7	142.93	204.64
Luty	2.86	2.73	160.1	184.39
Marzec	3.64	4.79	157.77	265.61
Kwiecień	5.82	12.21	193.63	240.87
Maj	8.62	13.43	207.46	142.01
Czerwiec	6.05	8.59	51.14	68.05
Lipiec	4.57	7.52	34.18	54.3
Sierpień	7.28	12.1	49.26	83.27
Wrzesień	9.89	13.53	200.2	255.5
Październik	4.44	7.28	163.4	227.21
Listopad	2.99	4.44	141.91	200.91
Grudzień	2.95	2.95	155.64	156.02

6.2.13 Wyznaczenie przedziałów ufności dla modelu całej sieci

Zagadnienie przedziałów ufności przedstawione w rozdziale pozwala na oszacowanie niedokładności modelu dla zadanego poziomu ufności $1 - \alpha$. Warunkiem wyznaczania przedsiałów ufności zgodnie z definicją 2.12 jest reprezentowanie przez błędy prognozy rozkładu normalnego. W związku z powyższym dla uzyskanych błędów prognozy na zbiorze testowym z podziałem na sezony wykonano test sprawdzający czy uzyskane błędy prognoz pochodzą z rozkładu normalnego [20]. Na podstawie uzyskanych wyników stwierdzono, że błędy prognoz nie reprezentują rozkładu normalnego dla żadnego z analizowanych sezonów. W następstwie do wykonanych testów określono alternatywną metodę wyznaczania przedziałów ufności nie wymagającą spełnienia warunku rozkładu normalnego dla błędów prognoz, tj.: wyznaczanie przedziałów ufności na podstawie wartości kwantyli dla błędów prognoz ze zbioru testowego. By uzyskać przedziały ufności dla poziomu ufności 90 proc. wyznaczono kwantyl

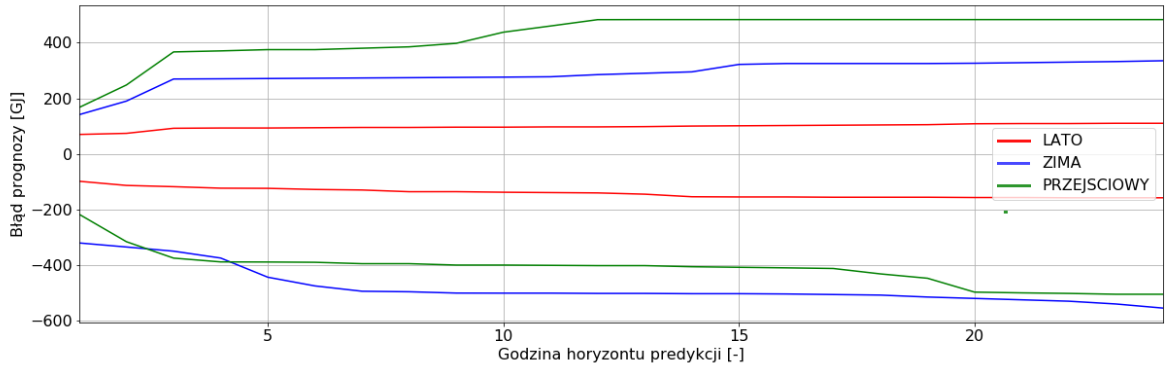
0,05 oraz 0,95 dla błędów prognoz dla danych testowych, dla których dane wejściowe stanowiły historyczne wartości danych pogodowych, uzyskane wartości dla modelu SSN bez autoregresji przedstawione są w tabeli 6.19.

Tabela 6.18: Model SSN z autoregresją: kwantyle błędów prognoz dla zbioru danych testowych w ujęciu godzinowym.

	sezon	**LATO**		**ZIMA**		**PRZEJŚCIOWY**	
godzina horyzontu predykcji	**kwantyl**	**0,05**	**0,95**	**0,05**	**0,95**	**0,05**	**0,95**
1		-98.6	69.9	-320.5	141.2	-217.7	167.2
2		-113.6	73.8	-335.0	189.9	-316.0	247.8
3		-118.0	92.0	-350.0	269.2	-375.0	366.7
4		-123.5	93.0	-374.7	270.0	-388.6	370.4
5		-124.0	93.0	-443.7	271.0	-389.0	375.0
6		-128.0	94.0	-475.0	272.0	-390.0	375.0
7		-130.0	95.0	-494.2	273.0	-395.0	380.0
8		-136.1	95.0	-496.0	274.0	-395.0	385.0
9		-136.0	96.0	-500.7	275.0	-400.0	397.6
10		-138.0	96.0	-501.0	276.0	-400.0	437.7
11		-139.2	97.0	-501.0	277.4	-401.0	460.0
12		-140.5	97.0	-502.0	285.0	-402.0	482.8
13		-145.0	98.0	-502.0	290.0	-402.0	483.0
14		-154.3	100.0	-503.0	295.0	-406.0	483.0
15		-155.0	101.0	-503.0	321.8	-408.0	483.0
16		-155.0	102.0	-504.0	325.0	-410.0	483.0
17		-156.0	103.0	-506.0	325.0	-412.3	483.0
18		-156.0	104.0	-508.0	325.0	-431.7	483.0
19		-156.0	105.0	-515.0	325.0	-447.4	483.0
20		-157.0	108.3	-520.0	326.0	-497.4	483.0
21		-157.0	109.0	-525.0	328.0	-500.0	483.0
22		-158.0	109.0	-530.0	330.0	-502.0	483.0
23		-158.0	110.0	-540.0	332.0	-505.0	483.0
24		-158.0	110.0	-554.8	335.0	-505.0	483.0

W przypadku modelu SSN z autoregresją błędy prognoz są analizowane indywidualnie dla każdej godziny horyzontu predykcji z uwagi na kumulowanie błędu predykcji, co zostało

przedstawione w sekcji 6.2.12. Wyznaczone wartości kwantyli błędów zostały skorygowane tak, aby reprezentowały tendencję wzrostową, ostateczne wartości przedstawione zostały w tabeli 6.18 oraz na wykresie 6.23 .



Rysunek 6.23: Kwantyle błędów prognoz dla modelu SSN z autoregresją

Tabela 6.19: Model SSN bez autoregresji: kwantyle błędów prognoz dla zbioru danych testowych.

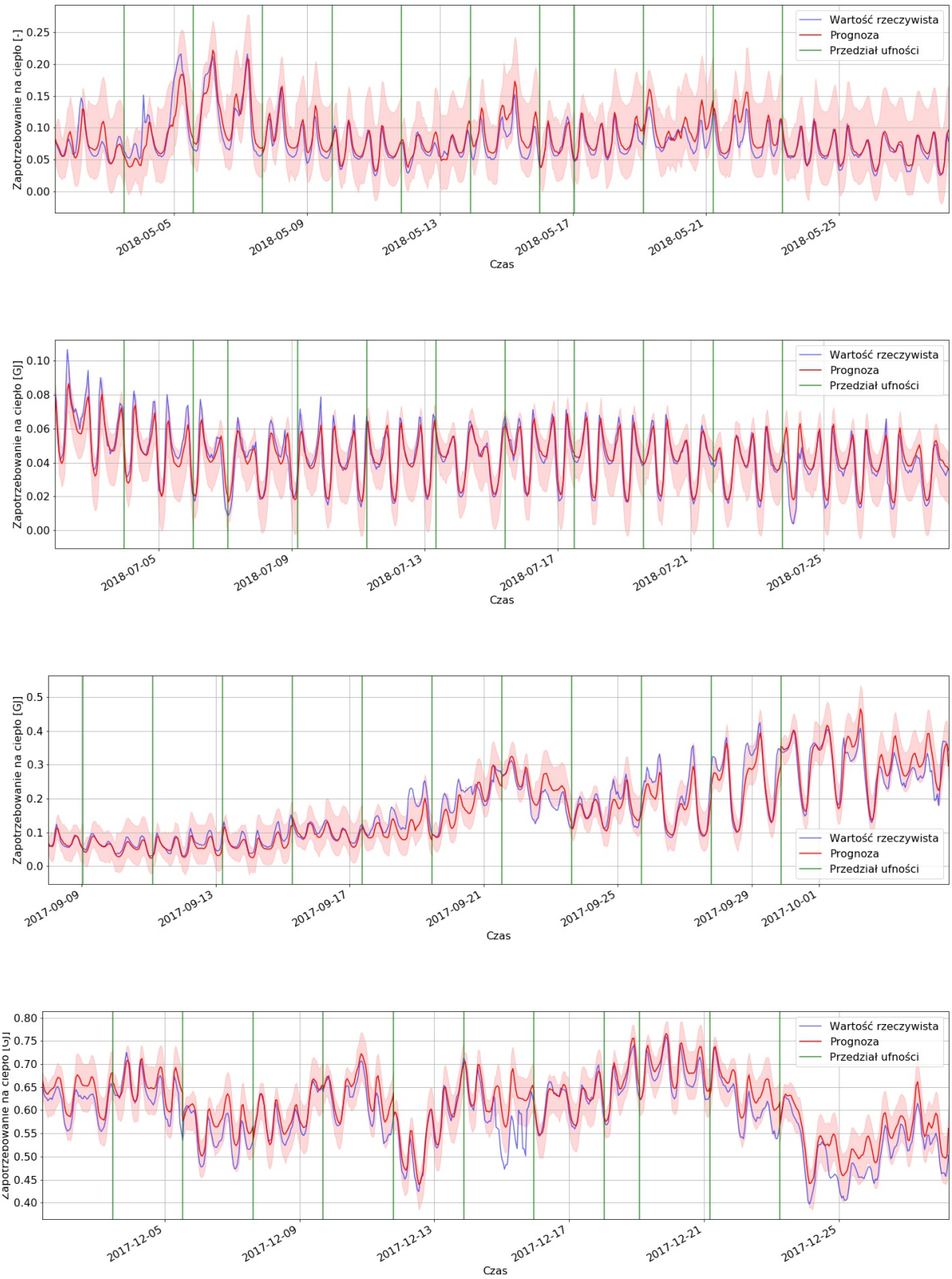
		SEZON		
		LATO	ZIMA	PRZEJŚCIOWY
KWANTYL	0,05	-160,22	-563,86	-548,69
	0,95	115,20	334,25	498,6

Dla nowych predykcji dla danego sezonu wartości przedziałów ufnosci wyznaczane są zgodnie z formułą 6.12. Wyznaczane w ten sposób przedziały ufności dla danego sezonu zawsze, niezależnie od wartości danych wejściowych, będą miały stałą szerokość, ponadto wyznaczane przedziały ufności nie uwzględniają niepewności związanych z precyzją prognozy pogody. Wyznaczone przedziały ufności spełnią swoje założenia, w przypadku gdy model utrzyma dokładność predykcji uzyskaną na zbiorze danych testowych. Na wykresach 6.24 przedstawiono prognozy zapotrzebowania na ciepło wraz z przedziałami ufności. Przedstawione dane pochodzą z szeregów czasowych przeskalowanych z wykorzystaniem parametrów skalowania WSC.

$$< y^* + q(sezon)_{0,05} : y^* + q(sezon)_{0,95} > \tag{6.12}$$

gdzie:

- y^* - wartość prognozy,
- $q(sezon)_{0,05}$ - kwantyl rzędu 0,05 błędów prognoz dla danych testowych dla wskazanego sezonu,
- $q(sezon)_{0,95}$ -kwantyl rzędu 0,95 błędów prognoz dla danych testowych dla wskazanego sezonu.



Rysunek 6.24: Wykresy znormalizowanej prognozy zapotrzebowania na ciepło wraz z przedziałami ufności.

6.2.14 Wyznaczenie najlepszego modleu dla obszarow sieci

By zwiększyć intuicję dyspozytorów sieci w obszarze poziomu zapotrzebowania na ciepło w różnych obszarach sieci, a tym samym poprawić kontrolę sieci opracowano dodatkowe modele zapotrzebowania na ciepło w obszarach sieci. Sieć ciepłownicza została geograficznie podzielona na 55 obszarów, które przedstawione zostały na rysunku 6.25. Wielkość obszaru, rozumiana jako liczba odbiorców przyporządkowana do danego obszaru jest zmienna i różnorodna, i została przedstawiona w tabeli 6.20.

Modele zapotrzebowania na ciepło dla obszarów sieci zostały opracowane z

Rysunek 6.25: Podział WSC na obszary.

wykorzystaniem doświadczenia z modelowania całkowitego zapotrzebowania na ciepło (sekcja 6.2.12), co oznacza, że zastosowano model autoregresyjnej regresji grzbietowej dla sezonu zimowego i sezonu letniego oraz modeli rozmytych dla sezonu przejściowego. Model sztucznej sieci neuronowej został pominięty ze względu na czasochłonne trenowanie modelu,

tj. 1 godzinę na model i porównywalne wyniki uzyskane we wstępnej analizie.

Tabela 6.20: Wielkość obszarów sieci.

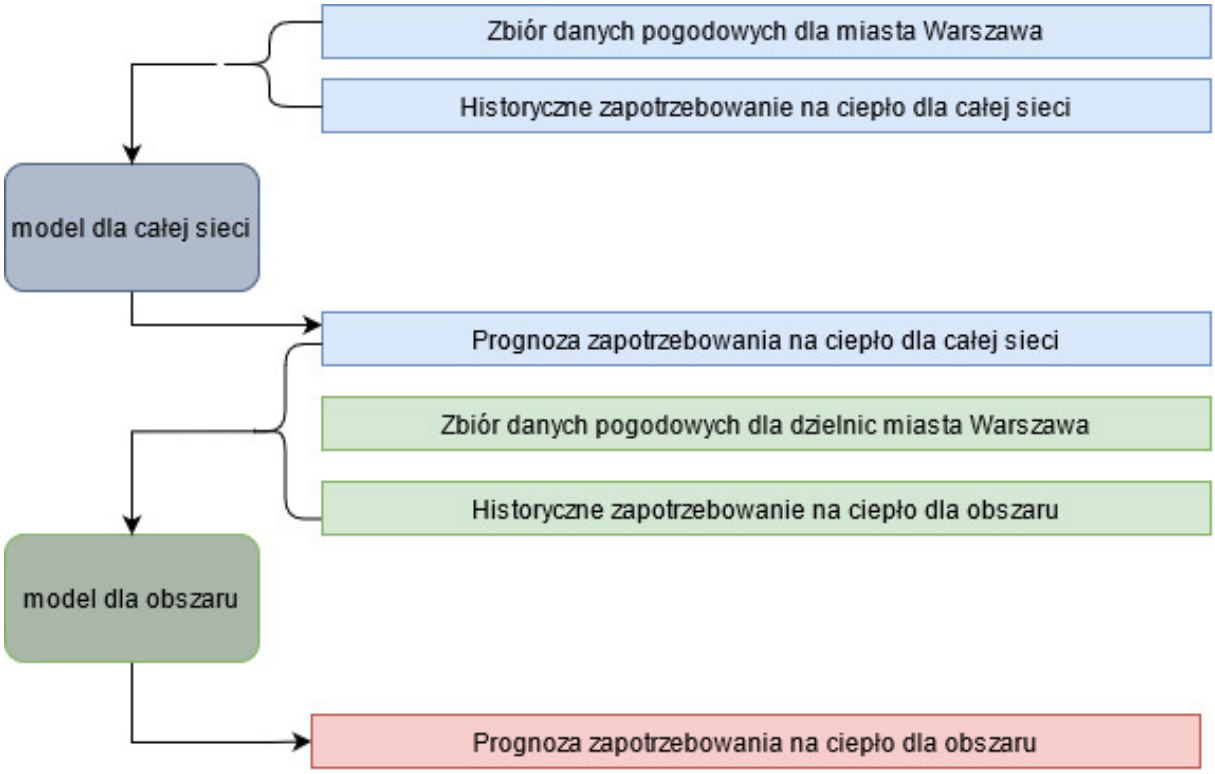
OBSZAR	ILOŚĆ WĘZŁÓW	OBSZAR	ILOŚĆ WĘZŁÓW	OBSZAR	ILOŚĆ WĘZŁÓW	OBSZAR	ILOŚĆ WĘZŁÓW
1	292	**16**	334	**31**	21	**46**	102
2	34	**17**	297	**32**	198	**47**	225
3	217	**18**	317	**33**	62	**48**	425
4	250	**19**	206	**34**	138	**49**	647
5	13	**20**	223	**35**	162	**50**	225
6	53	**22**	476	**36**	281	**51**	356
7	37	**23**	781	**37**	147	**52**	457
8	699	**24**	69	**38**	226	**53**	125
9	321	**25**	475	**39**	625	**54**	162
10	276	**26**	179	**40**	598	**55**	204
11	357	**27**	212	**42**	473		
12	322	**28**	379	**43**	701		
13	54	**29**	377	**44**	224		
14	157	**30**	416	**45**	362		
15	129						

Indywidualne modele dla obszarów składają się z trzech submodeli dla każdego sezonu, tj:

- model ridge regression dla sezonu letniego (model z autoregresją oraz model bez autoregresji);
- model ridge regression dla sezonu zimowego (model z autoregresją oraz model bez autoregresji);
- model dla sezonu przejściowego i model rozmycia wyników prognoz z wykorzystaniem modelu letniego oraz zimowego za pomocą funkcji rozmycia.

Dla sezonu przejściowego funkcja rozmycia jest identyczna z modelem dla całej sieci przedstawionym w rozdziale 6.2.12. Zmienne wejściowe modeli wybrierano idywidualnie dla każdego obszaru z wykorzystaniem selekcji wstępującej. Ponadto uwzględniono **dodatkową**

zmienną wejściową: wynik prognozy zapotrzebowania na ciepło dla całej sieci, która zawiera informację o referencyjnym poziomie zapotrzebowania na ciepło.



Rysunek 6.26: Metodologia wyznaczania zapotrzebowania na ciepło dla obszaru sieci.

Ideę wyznaczania progozy zapotrzebowania na ciepło dla obszarów przedstawiono na rysunku 6.26. W pierwszym etapie wyznaczana jest prognoza dla całej sieci na podstawie danych pogodowych dla miasta Warszawa, następnie uzyskany wynik prognozy oraz dane pogodowe dla dzielnicy przyporządkowanej do danego obszaru stanowią wartości wejściowe do modelu prognostycznego wyznaczającego prognozę zapotrzebowania na ciepło dla danego obszaru. Do trenowania modeli wykorzystano dane z okresu lipiec 2015–lipiec 2017. Finalnie liczba węzłów cieplnych przyporządkowanych do danego obszaru nie jest identyczna dla różnych obszarów i waha się od dziesiątek do setek węzłów cieplnych, ogólny wynik, tj. współczynnik MAPE dla każdego sezonu obliczono jako:

$$MAE_{sezon} = \frac{\sum_{i=1}^{liczba_{obszarw}} n_i * MAE_i}{\sum_{i=1}^{liczba_{obszarow}} n_i} \tag{6.13}$$

gdzie: $liczba_{obszarw}$ to liczba obszarów, dla których tworzone są modele; n to liczba podstacji przyporządkowanych do danego obszaru, MAE_i to MAE itego obszaru. Średnie wyniki dla sezonu przedstawione zostały w tabeli 6.21, natomiast wyniki dla zbioru testowego

indywidualnie dla każdego obszaru z podziałem na sezony przedstawione zostały w tabeli 6.22. Dla zobrazowania uzyskanych wyników histogram wyników dla parametru MAPE 1-24 godziny [%] z podziałem na sezony przedstawiony jest na rysunku 6.27

Wyniki dokładności modeli dla obszarów są zróżnicowane i w dużej mierze wynikają z wielkości węzłów cieplnych przyporzadkowanych do danego obszaru. Najlepsze dokładności prognoz uzyskiwane są dla sezonu zimowego z uwagi na większe, mocno skorelowane z temperaturą zewnętrzną zapotrzebowanie na ciepło. Natomiast w sezonie letnim jakość pomiarów opisana w rozdziale 6.2.2 wywiera dużo większy wpływ na możliwości prognostyczne modelu. Ponadto w przypadku małej liczby węzłów przyporządkowanych do obszarów indywidualne, nieskorelowane z temperaturą lub zmiennymi kalendarycznymi, zachowania klientów mają większy wpływ na chwilową wartość zapotrzebowania na ciepło dla danego obszaru. W sezonie przejściowycm występują okresy bardzo słabej jakości progozy, wynikające z uruchamiania wymienników c.o. które nie są bezpośrednio przewidywalne przez pogodę, a w konsekwencji przez model. Wykresy połączonych 24. godzinnych prognoz dla obszarów róznych wielkościowo (obszar 48: 425 odbioców, obszar 54: 162 odbiorców, obszar 18: 317, odbiorców ,obszar 7: 37 odbiorców) przedstawione zostały na wykresach: 6.29 - sezon letni, 6.28 - sezon zimowy oraz 6.30 - sezon przejściowy.

Tabela 6.21: Ogólne wyniki modeli dla obszarów sieci.

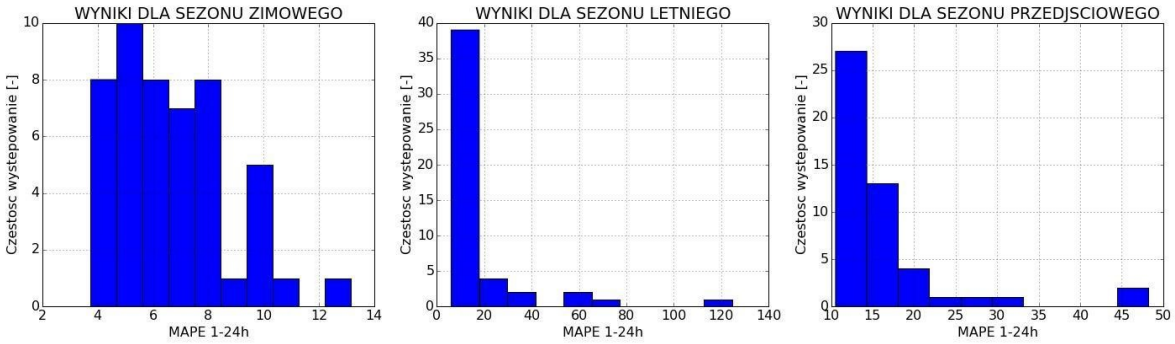
sezon	opis	1-24 godz.	25-72 godz.	1-24 godz.	25-72 godz.
		OBSZARY SIECI		INDYWIDUALNI ODBIORCY	
zimowy	MAE [GJ]	9.2	12.6	0.037	0.045
	średnia wartość [GJ]	136.7	136.7	0.05- 0.99	0.05- 0.99
przejściowy	MAE [GJ]	7.1	8.2	0.04	0.047
	średnia wartość [GJ]	71.8	71.8	0.1-0.49	0.1-0.49
letni	MAE [GJ]	1.63	1.84	0.008	0.01
	średnia wartość [GJ]	19.8	19.8	0 - 0.15	0 - 0.15

Tabela 6.22: Indywidualne wyniki modeli dla obszarów sieci.

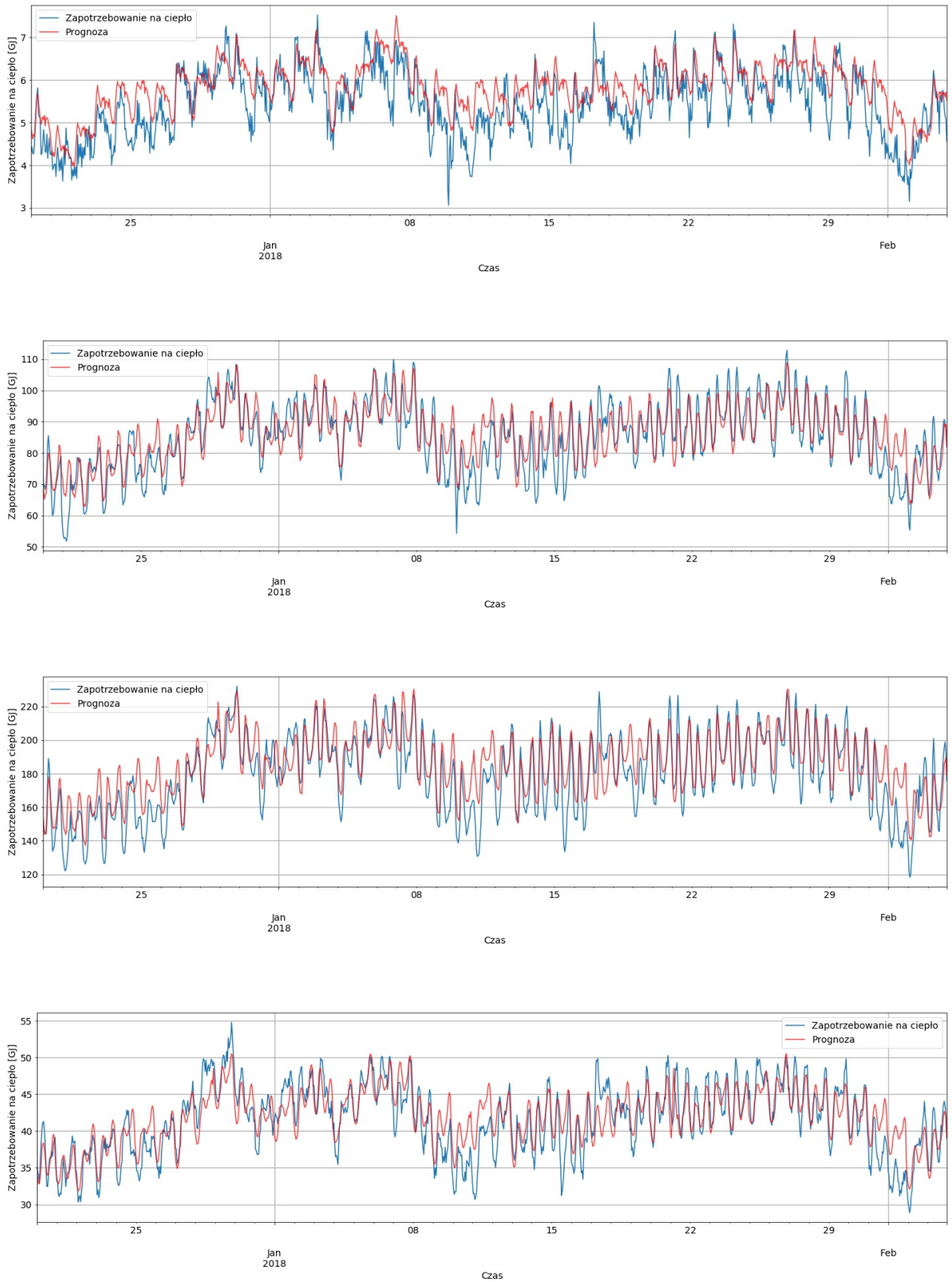
OBSZAR	SEZON	MAPE 1- 24h	MAPE 25-72 h
1	LATO	7.99	8.48
	PRZEJSCIOWY	12.75	14.87
	ZIMA	6.33	9.65
2	LATO	21.48	20.63
	PRZEJSCIOWY	19.4	21.86
	ZIMA	6.99	8.86
3	LATO	7.42	8.45
	PRZEJSCIOWY	14.96	14.68
	ZIMA	5.1	6.15
4	LATO	7.07	8.98
	PRZEJSCIOWY	13.13	16.29
	ZIMA	4.17	5.2
5	LATO	124.61	96.76
	PRZEJSCIOWY	48.18	46.02
	ZIMA	8.08	9.81
6	LATO	11.05	12.83
	PRZEJSCIOWY	24.25	18.42
	ZIMA	5.61	7.53
7	LATO	57.21	148.76
	PRZEJSCIOWY	27.51	18.32
	ZIMA	5.37	6.62
8	LATO	6.6	7.46
	PRZEJSCIOWY	12.6	14.92
	ZIMA	5.44	8.17
9	LATO	6.62	7.91
	PRZEJSCIOWY	15.47	19.14
	ZIMA	6.6	10.0
10	LATO	7.55	8.63
	PRZEJSCIOWY	11.31	12.62
	ZIMA	3.75	4.51
11	LATO	12.27	13.46
	PRZEJSCIOWY	12.69	13.88
	ZIMA	6.9	7.82
12	LATO	9.05	11.06
	PRZEJSCIOWY	19.15	15.16
	ZIMA	3.84	6.57
13	LATO	30.64	32.13
	PRZEJSCIOWY	17.39	20.34
	ZIMA	8.8	14.76
14	LATO	12.62	13.94
	PRZEJSCIOWY	12.5	13.88
	ZIMA	8.34	10.02
15	LATO	11.67	12.69
	PRZEJSCIOWY	13.32	14.42
	ZIMA	7.88	11.91
16	LATO	7.97	10.52
	PRZEJSCIOWY	12.02	13.68
	ZIMA	7.76	10.92
17	LATO	7.6	9.35
	PRZEJSCIOWY	13.43	13.83
	ZIMA	7.1	9.72
30	LATO	6.07	7.42
	PRZEJSCIOWY	15.27	23.96
	ZIMA	5.48	8.01
31	LATO	36.34	40.41
	PRZEJSCIOWY	32.15	51.21
	ZIMA	10.09	16.72
32	LATO	76.24	197.63
	PRZEJSCIOWY	17.54	17.44
	ZIMA	6.17	8.85
33	LATO	28.35	29.3
	PRZEJSCIOWY	20.14	23.49
	ZIMA	10.24	13.1
34	LATO	9.15	10.07
	PRZEJSCIOWY	12.09	13.82
	ZIMA	6.29	9.56
35	LATO	11.5	12.28
	PRZEJSCIOWY	10.63	12.43
	ZIMA	3.93	4.73
36	LATO	11.0	11.98
	PRZEJSCIOWY	11.76	13.46
	ZIMA	5.37	6.62
37	LATO	11.25	13.42
	PRZEJSCIOWY	11.92	12.63
	ZIMA	6.77	8.24
38	LATO	11.55	13.37
	PRZEJSCIOWY	11.85	14.58
	ZIMA	5.58	6.52
39	LATO	7.54	8.08
	PRZEJSCIOWY	11.18	11.27
	ZIMA	4.23	5.21
40	LATO	8.48	9.57
	PRZEJSCIOWY	10.53	11.25
	ZIMA	4.6	5.48
41	LATO	10.23	11.55
	PRZEJSCIOWY	11.27	12.16
	ZIMA	4.81	6.89
42	LATO	6.85	7.82
	PRZEJSCIOWY	11.25	12.35
	ZIMA	5.35	9.17
43	LATO	8.38	9.1
	PRZEJSCIOWY	12.45	11.8
	ZIMA	6.53	7.4
44	LATO	19.18	19.24
	PRZEJSCIOWY	17.93	23.95
	ZIMA	4.52	5.45
45	LATO	6.6	7.88
	PRZEJSCIOWY	14.55	18.24
	ZIMA	5.8	9.46
46	LATO	9.27	10.98
	PRZEJSCIOWY	13.76	15.98
	ZIMA	6.48	11.74

OBSZAR	SEZON	MAPE 1- 24h	MAPE 25-72 h
19	LATO	14.97	10.69
	PRZEJSCIOWY	13.06	14.61
	ZIMA	7.64	11.49
20	LATO	7.18	7.83
	PRZEJSCIOWY	15.89	16.96
	ZIMA	7.72	9.96
21	LATO	8.14	8.96
	PRZEJSCIOWY	13.49	13.47
	ZIMA	4.61	6.68
22	LATO	7.23	8.5
	PRZEJSCIOWY	15.94	18.33
	ZIMA	9.89	14.68
23	LATO	7.29	8.4
	PRZEJSCIOWY	15.32	16.4
	ZIMA	7.35	9.45
24	LATO	61.07	59.37
	PRZEJSCIOWY	47.56	50.29
	ZIMA	13.16	13.13
26	LATO	9.98	12.15
	PRZEJSCIOWY	10.94	13.16
	ZIMA	5.35	8.04
27	LATO	10.92	12.08
	PRZEJSCIOWY	12.45	12.92
	ZIMA	6.07	9.65

OBSZAR	SEZON	MAPE 1- 24h	MAPE 25-72 h
47	LATO	11.47	11.49
	PRZEJSCIOWY	13.41	14.96
	ZIMA	5.94	8.85
48	LATO	7.93	9.34
	PRZEJSCIOWY	14.39	16.42
	ZIMA	8.02	9.75
49	LATO	7.06	8.38
	PRZEJSCIOWY	14.39	15.63
	ZIMA	7.84	10.37
50	LATO	15.84	17.06
	PRZEJSCIOWY	14.35	14.63
	ZIMA	6.84	7.88
51	LATO	9.86	11.8
	PRZEJSCIOWY	13.16	14.4
	ZIMA	9.99	11.66
52	LATO	9.73	11.34
	PRZEJSCIOWY	13.11	13.45
	ZIMA	10.04	11.83
53	LATO	20.98	21.63
	PRZEJSCIOWY	18.29	19.83
	ZIMA	11.0	13.96
54	LATO	20,98	21,63
	PRZEJSCIOWY	18,29	19,83
	ZIMA	11	13,96
55	LATO	12,76	12,62
	PRZEJSCIOWY	15,34	16,96
	ZIMA	5,88	5,61

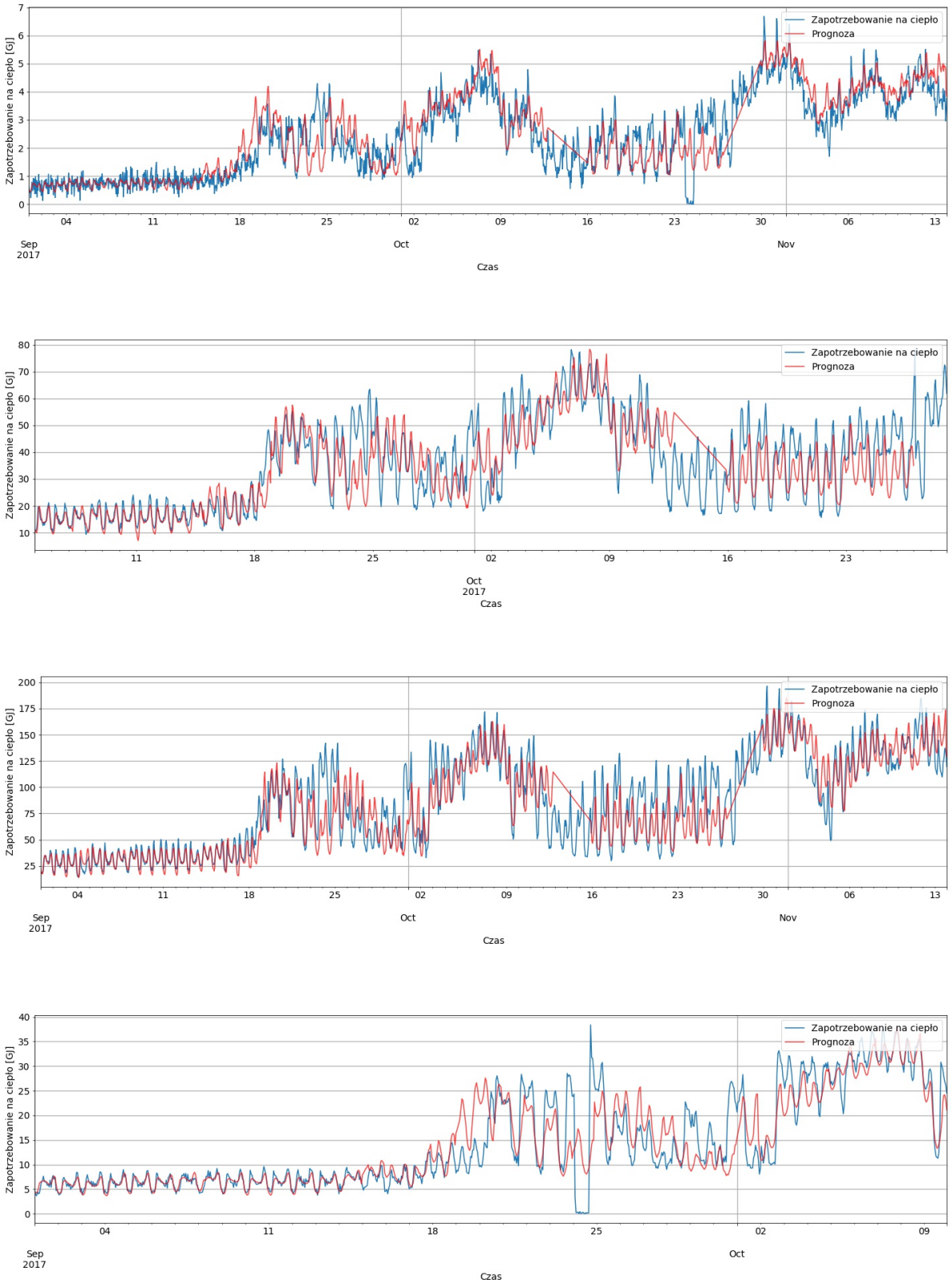


Rysunek 6.27: Histogram wyników dokładności prognozy dla obszarów.



Rysunek 6.28: Wykresy prognozy zapotrzebowania na ciepło dla wybranych obszarów dla sezonu zimowego.

Rysunek 6.29: Wykresy prognozy zapotrzebowania na ciepło dla wybranych obszarów dla sezonu letniego.

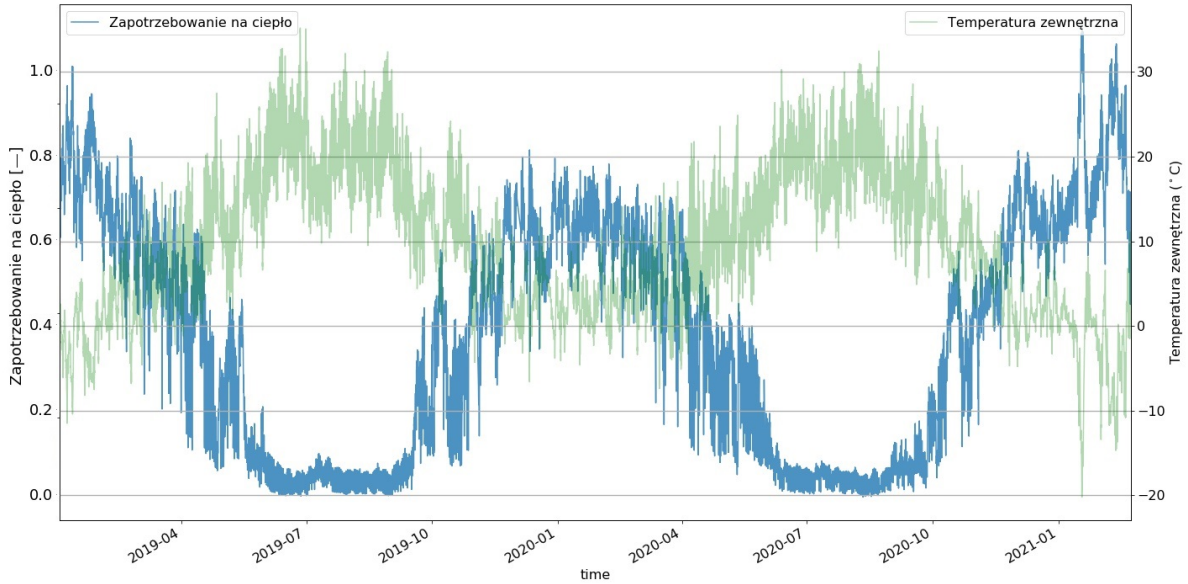


Rysunek 6.30: Wykresy prognozy zapotrzebowania na ciepło dla wybranych obszarów dla sezonu przejściowego.

Rozdział 7

Wyniki wdrożenia w systemie SWD

Modele prognostyczne zapotrzebowania na ciepło dla całej sieci oraz jej obszarów zostały wdrożone w systemie SWD pod koniec 2018. roku, w dalszej części rozdziału zaprezentowane zostaną wyniki, jakie uzyskał model w okresie styczeń 2019 – luty 2021. Wykres zapotrzebowania na ciepło przeskalowany z wykorzystaniem parametrów skalowania WSC oraz temperatura zewnętrzna dla analizowanego okresu przedstawiony został na rysunku 7.1.



Rysunek 7.1: Zapotrzebowanie na ciepło dla całej sieci w okresie styczeń 2019- luty 2021.

Okres ten charakteryzował się zróżnicowanymi warunkami pogodowymi: łagodną zimą w sezonie 2019/2020 oraz mocnymi utrzymującymi się mrozami w sezonie 2020/2021. Zauważalne jest, że wartości przeskalowanego zapotrzebowania na ciepło przekraczają wartość 1, co oznacza, że rzeczywista wartość zapotrzebowania na ciepło w przedstawianym

okresie wdrożenia przekroczyła maksymalną wartość występującą w danych treningowych, na podstawie których trenowano model oraz wyznaczono parametry skalowania WSC.

7.1 Wyznaczanie prognozy zapotrzebowania na ciepło w systemie SWD

Przedstawiona w rozdziale 6.2.12 kombinacja skierowanej SSN z autoregresyjnym wejściem i skierowanej SSN bez autoregresyjnego wejścia została zaimplementowana w systemie wspomagania decyzji dla warszawskiej sieci ciepłowniczej w celu wyznaczania prognozy zapotrzebowania na ciepło. Wyznaczenie prognozy zapotrzebowania na ciepło przebiega następująco:

1. Pobranie z bazy danych najnowszej dostępnej historii pomiarów zapotrzebowania na ciepło dla ciepłomierzy znajdujących w liście bazowej ciepłomierzy opisanej w sekcji 6.2.5.

2. Pobranie danych pogodowych:

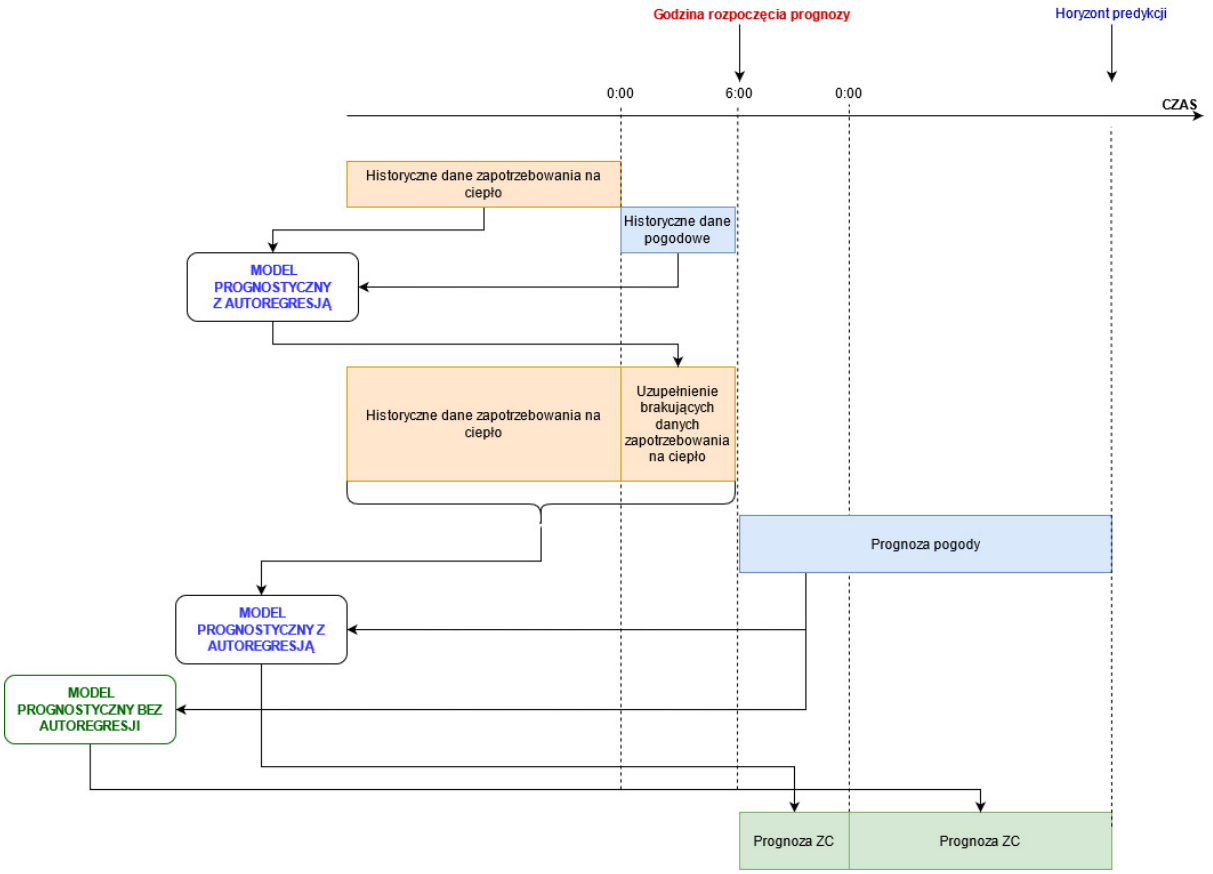
   - dla doby wstecz do obecnej chwili (godziny wyznaczania prognozy) pobierane są rzeczywiste, historyczne wartości danych pogodowych;
   - od obecnej chwili (godziny wyznaczania prognozy) do końca horyzontu predykcji pobierana jest prognoza pogody.

3. Walidacja danych z węzłów cieplnych wedle algorytmu opisanego w sekcji 6.2.2

4. Agregacja oraz uzupełnianie brakujacych pomiarów wedle metody opisanej w sekcji 6.2.7. W przypadku gdy liczba dostępnych pomiarów jest poniżej dopuszczalnego poziomu (75 proc.), wartości dla tychże kroków czasowych są uzupełnianie przez wyznaczenie prognozy zapotrzebowania na ciepło dla danej godziny z wykorzystaniem dostępnych danych historycznych.

5. Wyznaczenie prognozy oraz przedziałów ufności. Wynik prognozy zawiera:

   - prognozowaną wartość zapotrzebowania na ciepło;
   - przedziały ufności uwzględniające niepewność wyniku wynikającą z jakości modelu prognostycznego.



Rysunek 7.2: Wyznaczanie prognozy w systemie SWD.

Algorytm wyznaczania prognozy jest uruchamiany automatycznie o określonej godzinie lub aktywowany manualnie decyzją użytkownika systemu. Odczyty pomiarów zużycia ciepła są importowane do systemu raz dziennie o określonej godzinie, dlatego często przy obliczaniu prognozy występuje przerwa czasowa między ostatnimi dostępnymi pomiarami zużycia ciepła oraz punktem czasowym, w którym ustawiony jest początek prognozy. By pominąć ten problem, brakujące pomiary zużycia ciepła są wypełniane przez wyznaczanie prognozy zapotrzebowania na ciepło z horyzontem wstecznym dla brakującego okresu czasu z wykorzystaniem autoregresyjnej SSN z wejściem autoregresyjnym i historycznymi danymi pogodowymi jako danymi wejściowymi.

Przykładem takiej sytuacji jest przypadek, gdy odczyty pomiarów są importowane o północy, a prognozę do celów optymalizacji należy rozpocząć o 7. rano, w takim przypadku do obliczenia zużycia ciepła od północy do 7. rano stosuje się autoregresyjną SSN z autoregresją z wykorzystaniem historycznych danych pogodowych i w konsekwencji przez następne 17 godzin z wykorzystaniem prognozy pogody (oba horyzonty predykcji łącznie wynoszą 24 godziny), następnie model jest przełączany na SSN bez autoregresji. Schemat doboru źródeł

danych wejściowych i zastosowanych modeli prognostycznych przedstawiono na rysunku 7.2. Opisany schemat implementacji modelu okazuje się odporny na braki danych dotyczących zużycia ciepła i zapewnia najlepszą dostępną jakość prognozy.

7.2 Wyniki prognoz modelu dla całej sieci

W tej sekcji przedstawiona zostanie analiza dokładności prognoz zapotrzebowania na ciepło dla całej sieci wyznaczanych w systemie SWD dla okresu styczeń 2019– luty 2021. W tym okresie temperatura zewnętrzna dwukrotnie przekroczyła zakres, na podstawie którego trenowany był model, tj. 18 stycznia 2021 wystąpiła temperatura poniżej -18,6C, natomiast 26 i 30 czerwca 2019 roku wystąpiła temperatura powyżej 36,7 C. SWD wyznaczył **3320 prognoz zapotrzebowania na ciepło** w analizowanym okresie: 809 prognoz w okresie sezonu letniego, 1432 prognozy w okresie sezonu zimowego oraz 976 prognoz w okresie sezonu przejściowego.

Prognoza w systemie SWD wyznaczana jest automatycznie w momencie aktualizacji prognozy pogody, która występuje trzy razy w ciągu doby oraz na żądanie użytkowników systemów dla dowolnej godziny. Przedstawiona w tym rozdziale analiza uwzględnia pojedynczą prognozę dla każdego dnia. Jest to prognoza wyznaczana w godzinach porannych, która stanowi wartość wejściową do algorytmu optymalizacji.

Jakość modelu prognostycznego przedstawionego w sekcji 6.2.12 dotyczy prognoz wyznaczanych na podstawie rzeczywistych danych pogodowych, dlatego aby ocenić jakość prognoz modelu w okresie wdrożenia w stosunku do jakości modelu dla danch testowych wyznaczone zostały wsteczie prognozy zapotrzebowania na ciepło dla analizowanego okresu z wykrzystaniem rzeczywistych danych pogodowych.

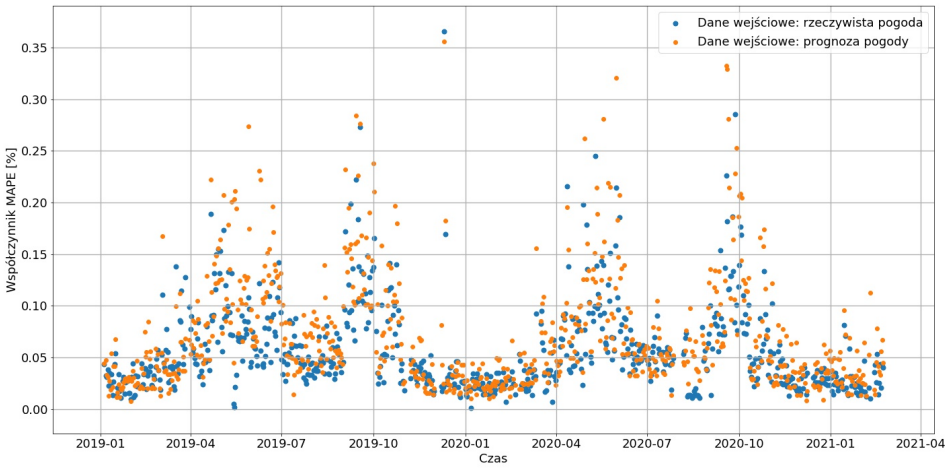
Prognozy z wykorzystaniem rzeczywistych danych pogodowych wyznaczone zostały tak, aby godzina rozpoczęcia prognozy pokrywała się z godziną rozpoczęcia automatycznych prognoz wyznaczonych z wykorzystaniem prognozy pogody. Uzyskane wyniki wyznaczone indywidualnie dla 2019 oraz 2020 roku, mierzone za pomocą współczynnika MAPE przedstawione zostały w tabeli 7.1.

Tabela 7.1: Współczynnik MAPE dla prognoz zapotrzebowania na ciepło po wdrożeniu w SWD.

		Dane wejściowe:	**rzeczywiste dane pogodowe**		**prognoza pogody**	
			1-24 godzina	**25-72 godzina**	**1-24 godzina**	**25-72 godzina**
WYNIKI	**ROK**	**SEZON**				
95% najlepszych dni	**2019**	**LATO**	6.18	7.78	8.91	10.4
		ZIMA	3.17	3.92	3.22	4.24
		PRZEJSCIOWY	7.04	8.13	8.88	11.61
	2020	**LATO**	5.34	7.01	6.41	7.91
		ZIMA	2.62	4	3.25	4.64
		PRZEJSCIOWY	7.62	8.62	9.43	10.35
wszystkie dni	**2019**	**LATO**	7.65	8.91	9.74	11.23
		ZIMA	3.91	4.7	3.95	4.69
		PRZEJSCIOWY	7.74	9.33	9.64	12.7
	2020	**LATO**	6.13	8.13	7.27	8.8
		ZIMA	2.87	4.21	3.54	5.03
		PRZEJSCIOWY	8.48	10.31	10.48	11.73

Wyniki są porównywalne z wynikami uzyskanymi dla zbioru danych testowych i potwierdzają jakość stworzonego modelu prognostycznego. Szczególnie istotne jest uzyskanie analogicznej (rok 2020) oraz lepszej (rok 2019) dokładności prognoz w okresie sezonu przejściowego tj., nie przekraczającej średniego błędu MAPE na poziomie 8,5 proc. dla pierwszych 24. godzin horyzontu predykcji oraz 10,5 proc. dla okresu od 25. do 72. godziny horyzontu predykcji. Wyniki dla sezonu zimowego również prezentują wysoką dokładność prognoz nie przekraczającą 4 proc. błędu MAPE dla pierwszych 24. godzin horyzontu predykcji oraz 5 proc. dla okresu od 25. do 72. godziny horyzontu predykcji. Dla sezonu letniego uzyskane wyniki są wyższe w porównaniu ze zbiorem danych testowych, gdzie dokładność prognoz dla pierwszych 24. godzin horyzontu predykcji jest gorsza w stosunku do wyników uzyskanych na zbiorze danych testowych o 1 punkt procentowy w 2020 roku oraz 2 punkty procentowe w roku 2019. Jednakże wynik ten nie wpływa istotnie na funkcjonowanie algorytmu optymalizacji oraz systemu SWD z uwagi na fakt, iż w sezonie letnim dostarczane jest wyłącznie ciepło na potrzeby ciepłej wody użytkowej.

Prognozy wyznaczane na podstawie prognozy pogody w porównaniu do wyników wyznaczonych na podstawie rzeczywistych wartości prezentują porównywalne wartości współczynnika MAPE dla sezonu zimowego, gdzie różnice w dokładności prognoz dla pierwszych 24. godzin horyzontu predykcji wynoszą 0,04 p.p. dla roku 2019 oraz 0,67 p.p. dla roku 2020. Większe różnice obserwowane są w sezonie letnim oraz przejściowym. W okresie sezonu letniego różnice wynoszą 2,09 p.p. (2019 rok) oraz 1,11 p.p. (2020 rok), natomiast w okresie sezonu przejściowego 2,90 p.p. (2019 rok) oraz 2,0 p.p. (2020 rok). Przedstawione różnice dokładności prognoz są zależne od jakości dostarczanych prognoz pogody, co oznacza, iż są poza obszarem wpływu samego modelu, w szczególności w przypadku sezonu przejściowego precyzyjna prognoza pogody może znacząco wpływać na jakość generowanych prognoz z uwagi na zależność pracy od temperatury zewnętrznej wymienników centralnego ogrzewania. Zaprezentowane różnice są akceptowalne i zadowalające z punktu widzenia dwuletniej pracy systemu SWD oraz algorytmu optymalizacji.

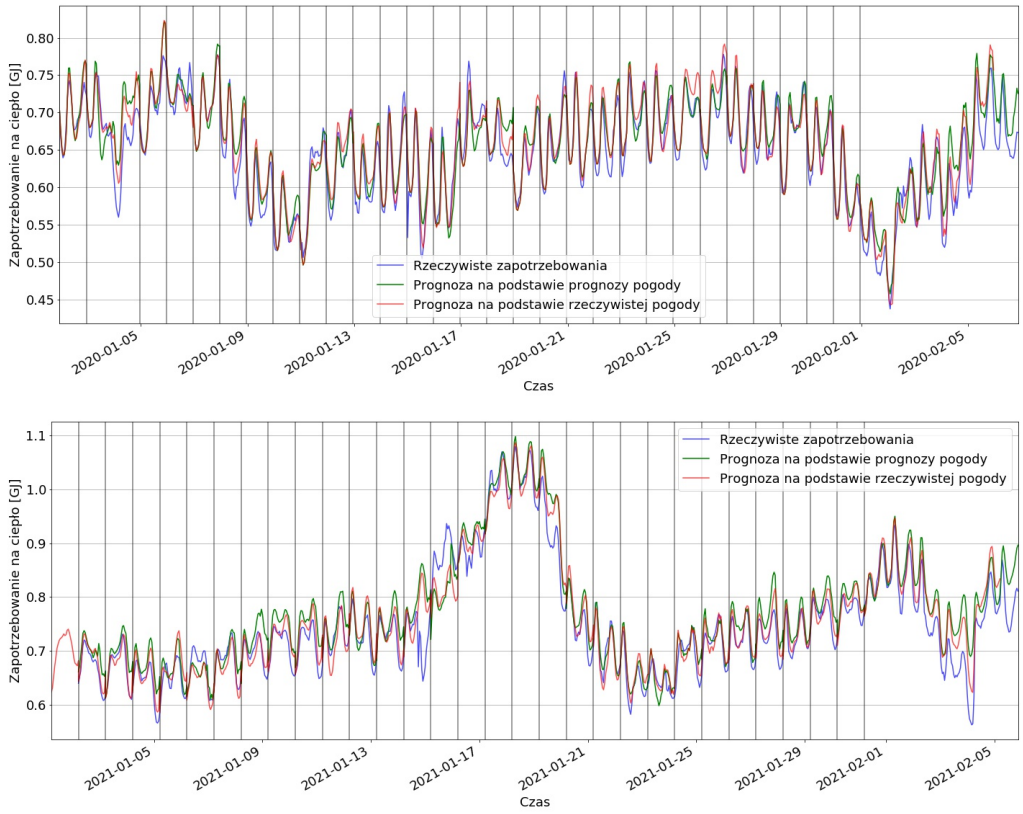


Rysunek 7.3: Wykres błędów prognoz. Styczeń 2019-Luty 2021.

Wartości współczynnika MAPE indywidualnie dla każdej wyznaczonej prognozy przedstawione zostały na wykresie 7.3 i pokazują, że obserwacje odstające nie są zależne od niepewności prognozy pogody, gdyż często występują dla tej samej daty generacji prognozy, niezależnie od danych wejściowych: historycznych wartości pogody lub prognozy pogody.

Wyniki dokładności prognoz z podziałem na lata i miesiące, przedstawione w tabeli 7.4, pokazują, że wzrost błędu prognozy w stosunku do zbioru testowego osiąga największe skoki dla miesięcy maj oraz wrzesień, są to miesiące zaliczane do sezonu przejściowego. Przebieg rzeczywistego zapotrzebowania na ciepło oraz prognozy dla tych miesięcy przedstawione

są na rysunku 7.6, pionowa linia oznacza moment rozpoczęcia prognozy. Zauważalne jest bardzo dobre odwzorowanie dynamiki zmian zapotrzebowania na ciepło, natomiast największe wyzwanie stanowi okres w drugiej połowie września, gdy uruchamiane są wymienniki ciepła oraz okres w drugiej połowie maja, gdy następuje całkowite wyłączenie centralnego ogrzewania. Pozostałe miesiące zaliczające się w większości dni do sezonu przejściowego, tj. kwiecień oraz październik przedstawione na wykresach celnie prognozują zarówno dynamikę zmian, jak i wartości zapotrzebowania na ciepło.

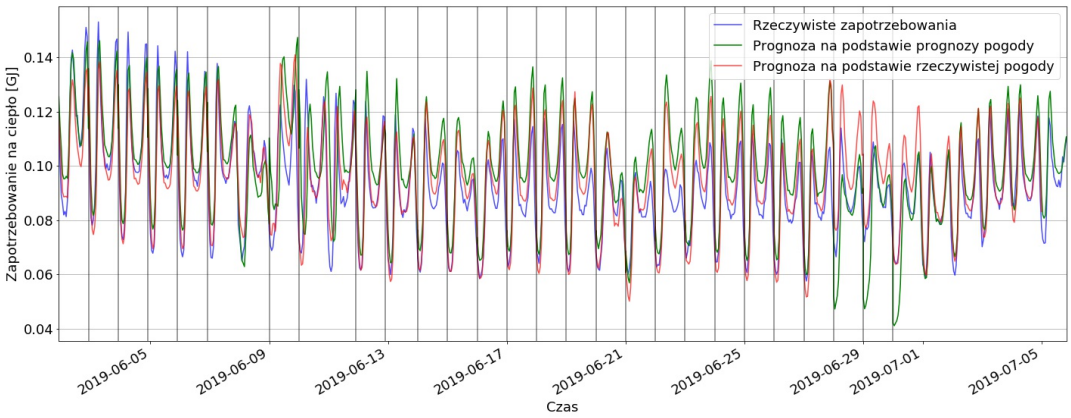


Rysunek 7.4: Prognoza zapotrzebowania na ciepło dla horyzontu 24 godzin. Styczeń 2020 oraz 2021.

Wysoką dokładność prognozy obserwuje się dla miesięcy zimowych: styczeń, luty, listopad oraz grudzień, gdzie współczynnik MAPE oscyluje w okolicach 3 proc.. Przebiegi rzeczywistych oraz prognozowanych wartości przedstawione zostały na wykresach 7.8 i 7.9. Uwagę zwracają wyniki prognoz wyznaczanych na podstawie prognozy pogody. Zauważalne jest pogorszenie prognozy w okresie styczeń–luty 2021 roku (MAPE 4,09 proc.) w stosunku do wyników uzyskiwanych dla tych samych miesięcy z lat poprzednich, zarówno w porównaniu ze zbiorem danych testowych, jak i wyników z okresu wdrożenia dla 2019 oraz 2020 roku.

Prognozy dla stycznia 2020 oraz 2021 roku przedstawiony został na wykresie 7.4, który zawiera 24 godzinne prognozy oraz rzeczywiste zapotrzebowanie na ciepło, gdzie pionowe zielone linie oznaczają moment rozpoczęcia prognozy. Analizując wykresy, zauważalne jest, że model prognozował znacznie większe zapotrzebowanie na ciepło niż występowało w rzeczywistości. Należy tutaj zwrócić uwagę, że profil zapotrzebowania na ciepło różni się od profilu z tego samego miesiąca lat poprzednich: wyraźne dwa wzrosty i spadki zapotrzebowania w cyklu dobowym zastąpione zostały jednym znacznym spadkiem oraz drugim o mniejszej wartości. Taka zmiana profilu zmiennej prognozowanej, odzwierciedlająca zmianę zachowania odbiorców ciepła, może być przyczyną pogorszonych wyników prognoz, w stosunku do precyzyjnej prognozy (MAPE poniżej 4 proc.) w latach poprzednich. Przedstawiona zmiana zachowania odbiorców w 2021 roku może wynikać z występujących w tym czasie restrykcji związanych z pandemią COVID-19 [19] takich jak: nauczanie zdalne, brak wydarzeń masowych, zamknięcie restauracji, zintensyfikowana praca zdalna itp.

Sezon letni dla miesięcy wakacyjnych, tj. lipiec–sierpień uzyskuje stabilne wyniki, 24.

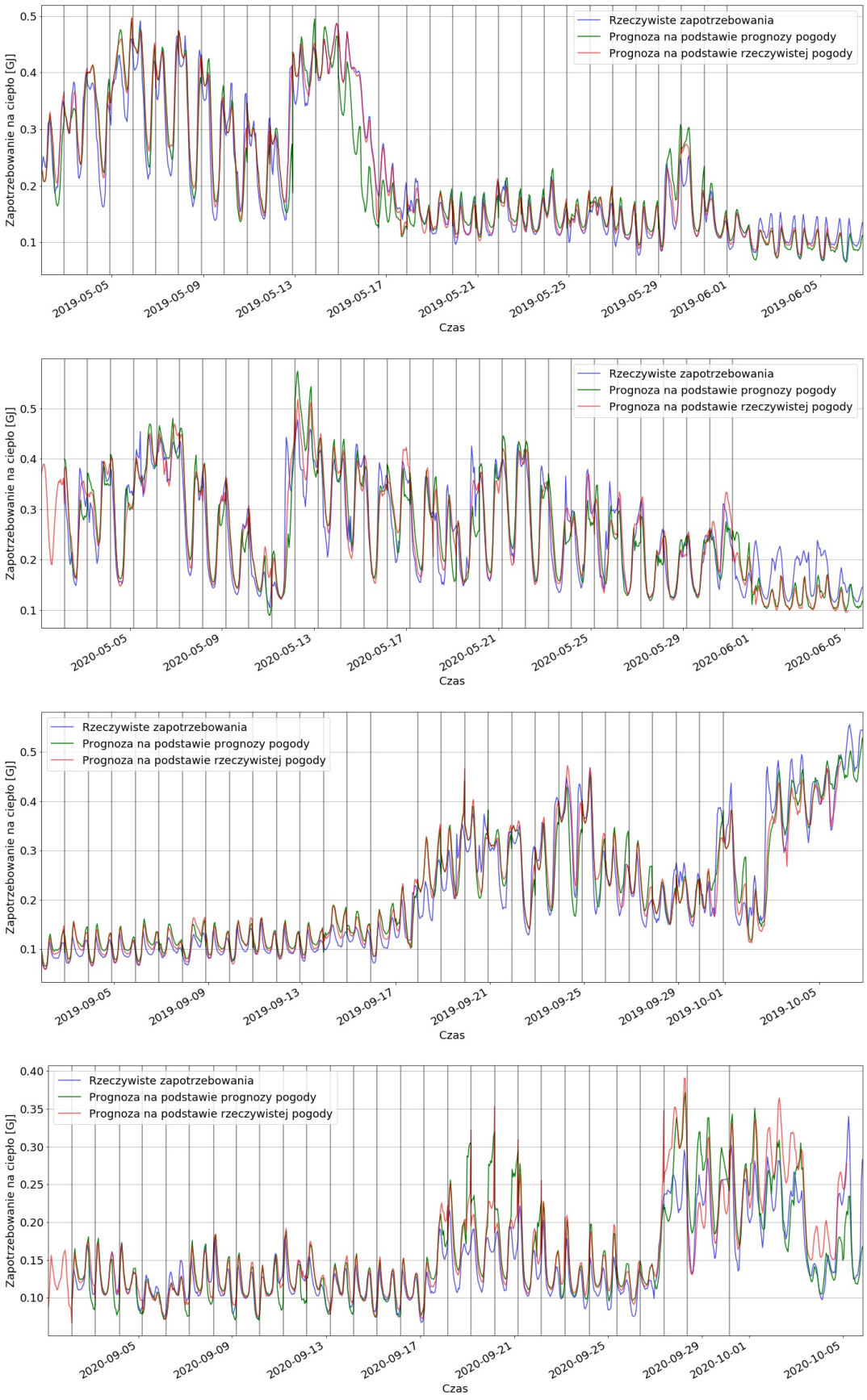


Rysunek 7.5: Prognoza zapotrzebowania na ciepło dla horyzontu 24 godzin. Czerwiec 2019 rok.

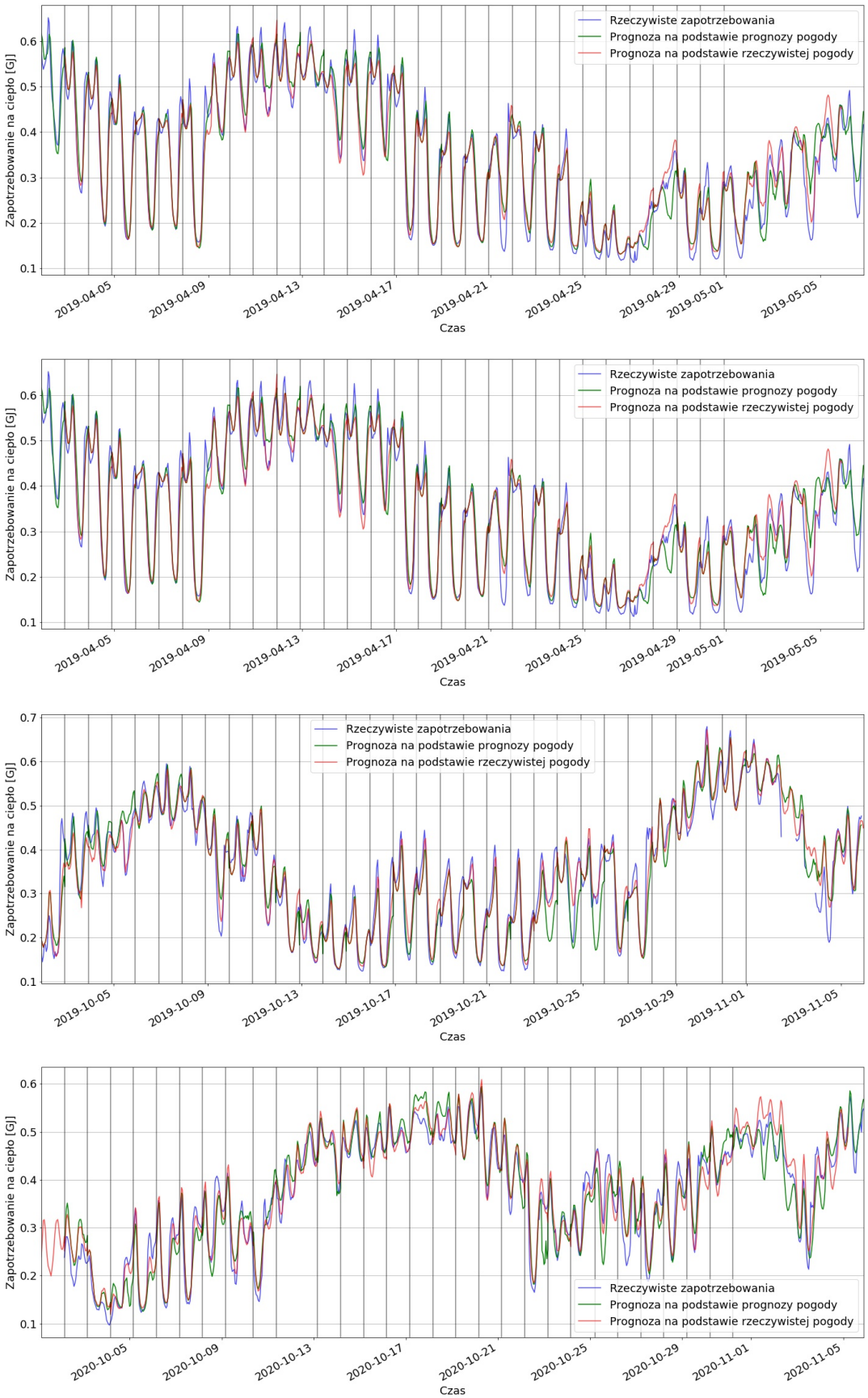
godzinna prognoza, gdzie współczynnik MAPE nie przekracza 5 proc. przedstawione zostały na wykresie 7.10. W sezonie letnim wyróżnia się miesiąc czerwiec 2019 roku, gdzie średni błąd prognozy MAPE przekracza 10 proc.. 24.godzinne prognozy przedstawione są na wykresie **??**. Zauważalny jest znaczny spadek poziomu zapotrzebowania na ciepło oraz obniżona wartość szczytów zapotrzebowania.

Tabela 7.2: Wyniki prognoz zapotrzebowania na ciepło po wdrożeniu w SWD z podziałem na lata i miesiące

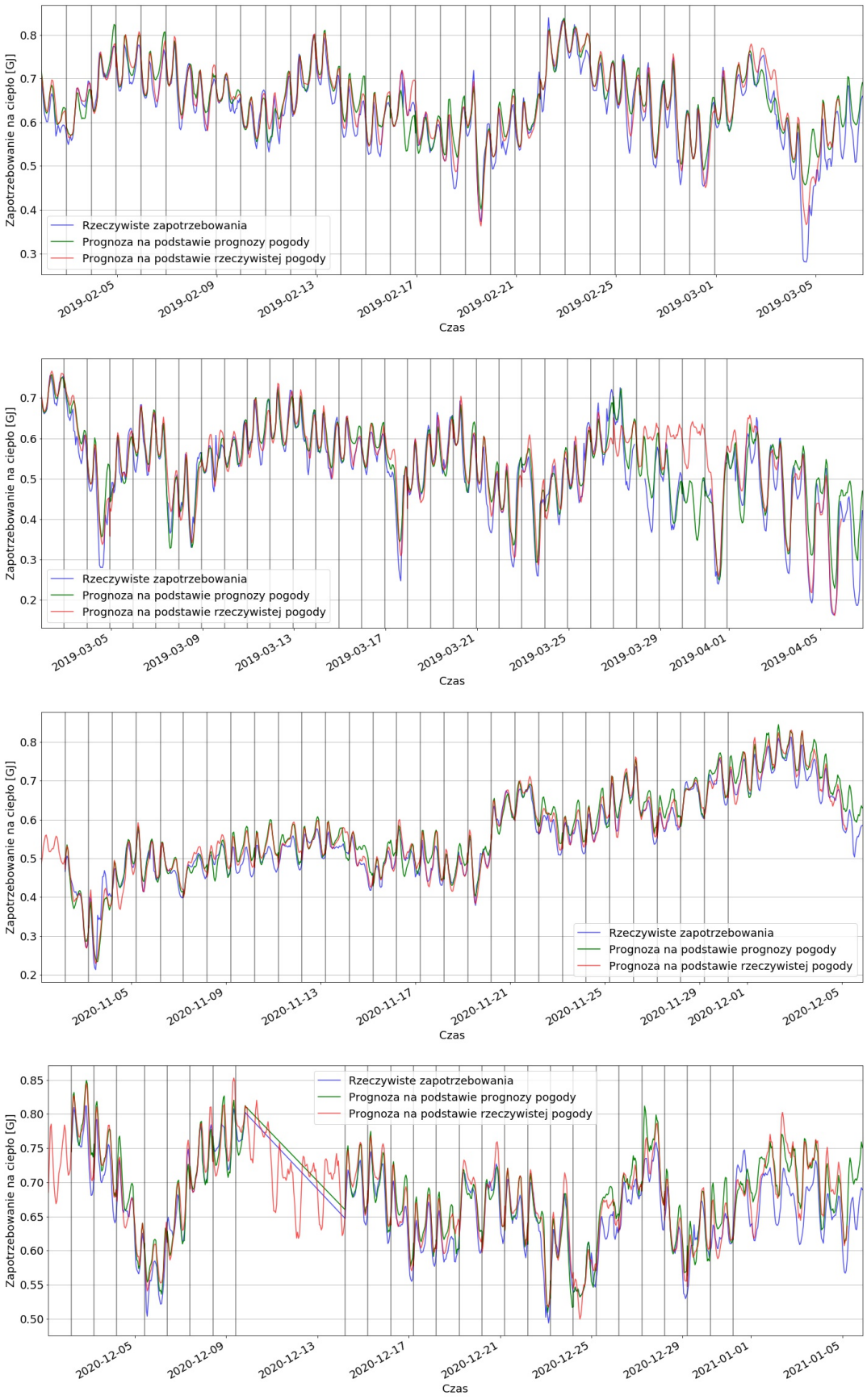
		MAPE [%]			
		Dane wejściowe: rzeczywista pogoda		Dane wejściowe: prognoza pogody	
		1-24 godzina	25-72 godzina	1-24 godzina	25-72 godzina
2019	1	2.63	2.7	2.85	2.97
	2	3.2	3.65	3.74	4.39
	3	6.01	7.88	4.98	6.32
	4	7.83	8.77	9.26	12.05
	5	8.77	14.33	12.68	17.99
	6	7.64	10.41	12.35	11.44
	7	4.95	7.16	5.87	11.7
	8	4.73	5.69	6.7	8.27
	9	15.04	12.85	15.59	14.16
	10	7.1	7.34	9.13	11.67
	11	3.35	3.59	3.94	5.99
	12	4.63	5.58	4.81	5.28
2020	1	2.17	4.66	2.44	4.69
	2	2.7	4.03	3.1	4.47
	3	3.91	4.63	5.43	6.86
	4	7.22	7.81	8.37	11.11
	5	11.31	12.19	13.17	13.64
	6	6.53	8.88	7.64	8.77
	7	4.97	6.07	5.03	6.9
	8	3.42	5.87	5.47	7.04
	9	11.04	17.89	13.73	16.05
	10	7.18	7.48	9.51	9.77
	11	3.39	4.05	4.31	5.13
	12	2.88	3.75	3.3	4.26
2021	1	2.76	4.4	3.65	5.43
	2	2.89	3.81	4.09	4.95



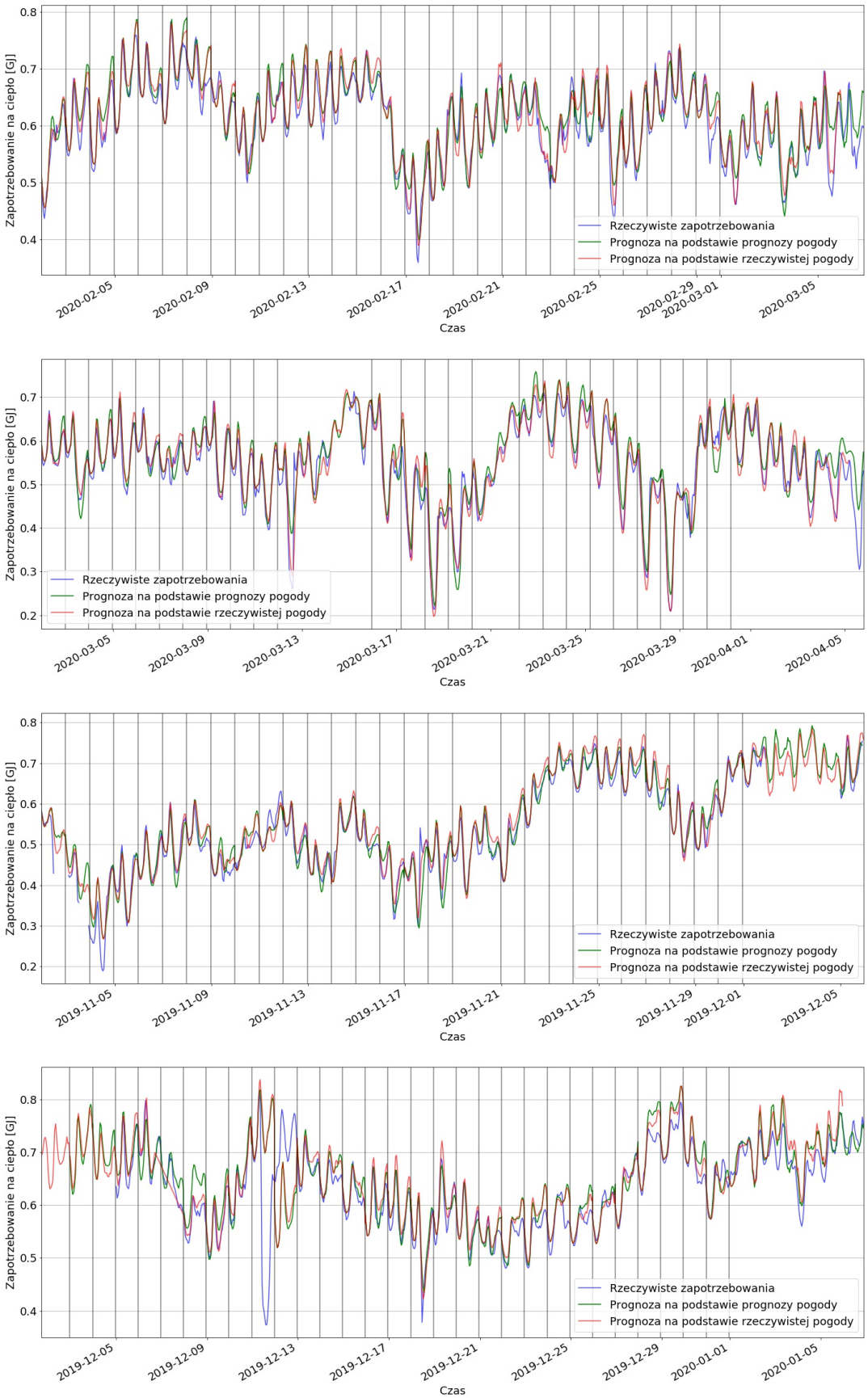
Rysunek 7.6: Prognoza zapotrzebowania na ciepło dla horyzontu 24 godzin. Maj– wrzesień 2019 oraz 2020 roku.



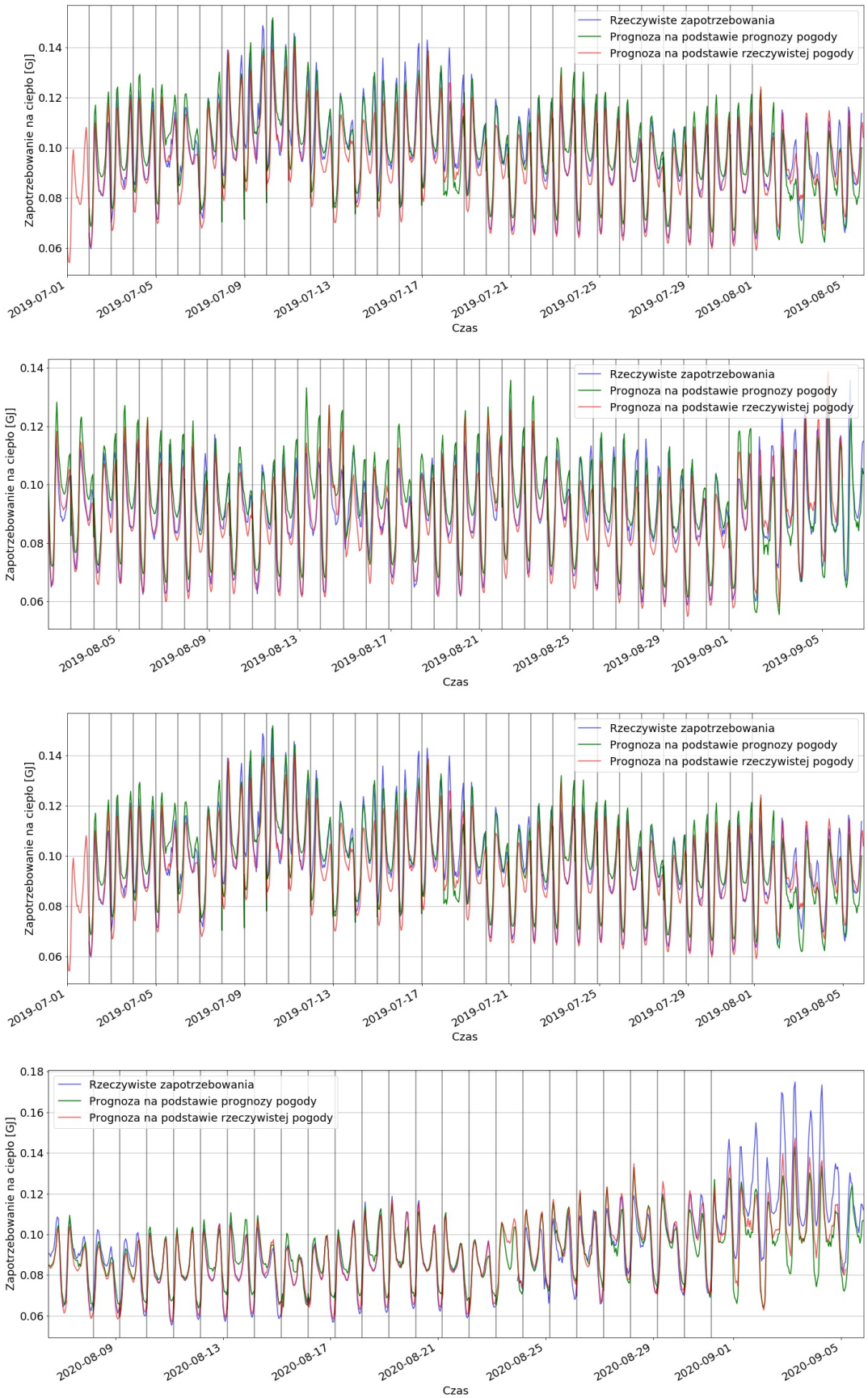
Rysunek 7.7: Prognoza zapotrzebowania na ciepło dla horyzontu 24 godzin. Kwiecień–październik 2019 oraz 2020 roku.



Rysunek 7.8: Prognoza zapotrzebowania na ciepło dla horyzontu 24 godzin. Luty, marzec, listopad, grudzień 2019 roku.



Rysunek 7.9: Prognoza zapotrzebowania na ciepło dla horyzontu 24 godzin. Luty, marzec, listopad, grudzień 2020 roku.



Rysunek 7.10: Prognoza zapotrzebowania na ciepło dla horyzontu 24 godzin. Lipiec, sierpień 2019 oraz 2020 roku.

7.3 Efektywność przedziałów ufności

Efektywność przedziałów ufnosci wyznaczanych w rozdziale 6.2.13 oceniano na podstawie parametru prawdopodobieństwa pokrycia przedziału prognozy (ang. Prediction Interval Coverage Probability) (PICP), który określa procent czasu, w którym przedział ufności zawierał rzeczywistą wartość prognozy oraz jest wyznaczany zgodnie z formułą 7.1.

$$PICP_{d(x),g(x))} = \frac{1}{N}\sum_{i=1}^{N} h_i \;\; gdzie\ h_i = \begin{cases} 1 & jezeli\ d(x) \leq y_i \leq g(x) \\ 0 & w\ przeciwnym\ razie \end{cases} \tag{7.1}$$

gdzie:

- $d(x), g(x)$ - dolna oraz górna granica przedziału ufności,
- y_i - wartość rzeczywista zmiennej prognozowanej,
- N - wielkość analizowanego zbioru danych.

Wartości PICP dla prognoz wyznaczonych na podstawie rzeczywistych danych pogodowych przedstawiono w tabeli 7.3. W miejscu tym należy przypomnieć, że prognozy dla pierwszych 24. godzin horyzontu predykcji wyznaczane są na podstawie modelu SSN z autoregresją natomiast kolejne godziny od 25. do 72. na podstawie SSN bez autoregresji.

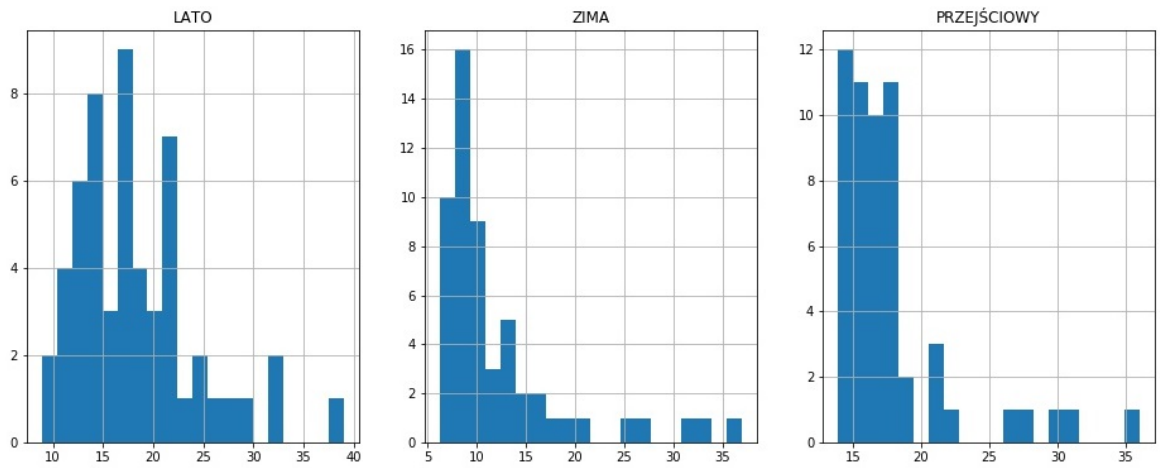
Tabela 7.3: Efektywność przedziałów ufności.

		godziny horyzontu predykcji	
		1-24	**25-72**
		PICP [%]	**PICP [%]**
2019	**LATO**	85.8	82
	ZIMA	91.8	90.5
	PRZEJSCIOWY	88.2	89
2020	**LATO**	84.3	79.3
	ZIMA	96.6	95.5
	PRZEJSCIOWY	88.4	89.4

Wartość oczekiwana parametru PICP wynosi 90proc., ponieważ dla takiego poziomu ufności wyznaczone zostały wartości przedziałów. Dla sezonu zimowego uzyskano zakładany poziom ufności przedziałów, najniższa jego wartość wynosi 90,5 proc. dla roku 2019 oraz obszaru od 25. do 72. godziny horyzontu predykcji. Sezon przejściowy uzyskuje wartości maksymalnie o 1,8 p.p. niższe od oczekiwanej. Najmniejszą efektywnością przedziałów ufności charakteryzuje się sezon letni, gdzie PICP plasuje się średnio w obszarze 85 proc., co daje wynik o 5 p.p niższy od oczekiwanego. Uzyskana efektywność przedziałów ufności stanowi potwierdzenie dobrej jakości modelu predykcyjnego, w szczególności w okresie sezonu zimowego oraz przejściowego.

7.4 Model dla obszarów sieci

Prognozy dla obszarów w systemie SWD wyznaczane były analogicznie jak prognoza dla całej sieci codziennie w godzinach porannych. Wyniki prognoz z podziałem na sezony dla okresu lipiec 2019– luty 2021 przedstawione zostały w tabeli 7.4, natomiast rozkład wyników prognoz przedstawiony został na wykresie 7.11.



Rysunek 7.11: Wykres rozkładów wyników prognoz dla obszarów z podziałem na sezony.

Najlepsza dokładność pogody w sezonie zimowym występuje dla obszaru o numerze 36, gdzie MAPE wynosi 6,3 proc., ponad połowa obszarów (36 obszarów z 55) uzyskuje dla sezonu zimowego średnią dokładność prognozy w horyzoncie 24. godzin poniżej 10 proc., jednocześnie cztery obszary (obszar 5,33,44,55) uzyskują dokładność prognozy powyżej 25 proc. .

W sezonie letnim, najlepsze wyniki otrzymywane są dla obszaru o numerze 30, gdzie MAPE wynosi 8.9 proc., mediana współczynnika MAPE wynosi natomiast 17,30 proc., dla 7 obszarów (Obszar 2, 7, 13, 26, 28, 33, 44) uzyskuje wyniki powyżej 25 proc.. Wyniki dla sezonu przejściowego, z uwagi na wykorzystywany model rozmyty stanowią pochodną wyników dla sezonu letniego oraz zimowego. Najlepsze wyniki prognozy uzyskane zostały dla obszaru 48 z wartością MAPE 13,9 proc., mediana wartości parametru MAPE wynosi 16,3 proc., a 5 obszarów (obszar 2, 5, 31, 33, 34) prezentuje wynik MAPE powyżej 25 proc..

Tabela 7.4: Wyniki prognoz zapotrzebowania na ciepło po wdrożeniu w SWD z podziałem na lata i miesiące.

	MAPE		
	LATO	ZIMA	PRZEJŚCIOWY
1	12.8	7.6	14.0
2	25.1	15.9	27.4
3	14.8	12.5	17.8
4	17.3	8.5	15.2
5	12.8	36.9	36.0
6	16.5	9.4	18.8
7	31.5	9.8	21.0
8	11.1	8.0	14.0
9	12.6	8.1	14.3
10	21.5	16.7	16.4
11	9.2	9.7	14.6
12	15.1	13.1	16.8
13	29.6	13.9	22.6
14	20.9	8.9	15.4
15	21.2	8.0	15.4
16	17.9	9.6	15.4
17	12.7	10.0	16.3
18	13.1	9.9	16.2
19	16.1	11.6	17.9
20	12.0	11.9	17.3

	MAPE		
	LATO	ZIMA	PRZEJŚCIOWO
21	14.5	7.2	14.4
22	10.9	7.0	14.6
23	16.7	9.7	17.4
24	14.0	7.1	17.4
25	16.6	8.6	14.9
26	31.7	18.9	17.3
27	22.0	8.4	16.8
28	25.6	12.5	19.1
29	22.2	7.4	15.6
30	8.9	8.8	14.1
31	22.3	14.9	30.4
32	11.9	8.2	15.1
33	39.0	26.6	30.8
34	23.8	6.8	16.2
35	17.9	7.6	17.1
36	13.6	6.3	15.8
37	19.2	8.7	16.3
38	17.9	18.3	16.8
39	18.7	9.0	18.1
40	18.0	8.3	14.1

	MAPE		
	LATO	ZIMA	PRZEJŚCIOWO
41	20.3	21.5	21.4
42	15.1	8.7	14.7
43	17.2	9.4	16.3
44	28.0	32.1	27.1
45	14.6	9.2	15.5
46	22.4	14.7	17.5
47	19.3	9.6	17.6
48	14.1	7.5	13.9
49	14.4	8.0	15.8
50	19.9	9.1	15.7
51	17.7	7.2	14.2
52	21.2	12.9	15.6
53	24.0	24.9	21.5
54	10.5	12.0	17.5
55	14.1	33.3	18.1

Uzyskane wyniki w świetle modelu zapotrzebowania na ciepło dla całej sieci są satysfakcjonujące, wybrane obszary, takie jak obszar 2, 7, 33 oraz 44 uzyskujące skrajne wyniki wymagają ponownej rekalibracji z użyciem większej ilości danych treningowych, gdyż

możliwe jest, że nastąpiła istotna zmiana zachowania odbiorców wpływająca na pogorszenie predykcji modelu, w szczególności w przypadku obszarów 2, 7 oraz 33 dla których liczba węzłów cieplnych przyporządkowanych do danego obszaru jest mniejsza niż 70.

7.5 Ogólne wnioski z wdrożenia

Wyniki uzyskane w trakcie ponad 2-letniej pracy, której wyniki przedstawiono w niniejszej pracy, nad systemem SPROG pozwalają na sformułowanie następujących wniosków:

- stworzony model prognostyczny dla zadanego horyzontu predykcji jest w stanie generować prognozy zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej;
- model po wdrożeniu uzyskuje dokładność prognoz porównywalną z wynikami uzyskanymi dla danych ze zbioru testowego, tj. błąd MAPE poniżej 4 proc. dla sezonu zimowego, poniżej 8,5 proc. dla sezonu przejściowego oraz poniżej 7,65 proc. dla sezonu letniego;
- dokładność predykcji modelu zapotrzebowania na ciepło dla całej sieci ciepłowniczej dla okresu wdrożenia dla pierwszych 24 godzin horyzontu predykcji jest istotnie lepsza od wyników prezentowanych w literaturze światowej, gdzie dla okresu grzewczego prezentowana dokładność predykcji liczona współczynnikiem MAPE wynosi od 4,76 proc. do 11,7 proc., natomiast prezentowany w niniejszej pracy model uzyskuje dla sezonu grzewczego wyniki poniżej 4 proc. błędu MAPE. Z uwagi na brak wyników dla pozostałych okresów roku, tj. sezonu letniego oraz przejściowego porównanie wyników nie jest możliwe;
- stworzone dodatkowe modele prognostyczne są w stanie generować prognozy zapotrzebowania na ciepło dla obszarów warszawskiej sieci ciepłowniczej;
- generowane prognozy zapotrzebowania na ciepło dla obszarów WSC dla większości obszarów sieci osiągają akceptowalną dokładność predykcji.

Rozdział 8

Podsumowanie

W niniejszej pracy przedstawiono system generujący prognozy zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej. System prognozuje zapotrzebowanie na ciepło dla całej sieci oraz jej zdefiniowanych obszarów dla zadanego horyzontu predykcji. Wynik prognozy zawiera również przedziały ufności prognozy dla poziomu ufności równego 90 proc.. Prognozy generowane są na podstawie modeli prognostycznych, które stworzono z wykorzystaniem technik uczenia maszynowego. W rozprawie przedstawiono metody prognozowania szeregów czasowych oraz metodologię budowania systemów opartych o techniki uczenia maszynowego.

Prezentowany system prognostyczny, w przeciwieństwie do rozwiązań opisywanych w światowej literaturze naukowej, bazuje na danych pomiarowych indywidualnych odbiorców ciepła, tj. danych pomiarowych z węzłów cieplnych przyłączonych do warszawskiej sieci ciepłowniczej. Rozwiązanie to pozwala na skuteczniejsze planowanie oraz eksploatację systemu ciepłowniczego, oraz transformację w kierunku systemu ciepłowniczego czwartej generacji. Opracowano algorytmy pozwalające na wykorzystanie danych pomiarowych: algorytm walidacji danych, algorytm agregacji danych i algorytm dynamicznego uzupełniania brakujących danych. Z uwagi na fakt, iż aktualna liczba użytkowników systemu jest zmienna w czasie przedstawiono koncepcję listy bazowej, która pozwala na operowanie stałą liczbą odbiorców w procesie generowania prognoz zapotrzebowania na ciepło i jednocześnie umożliwia skalowanie wyników prognoz na zmienną liczbę aktualnych użytkowników systemu. Na tej podstawie uważa się, że **potwierdzono tezę główną pracy i możliwe jest stworzenie systemu realizującego prognozy zapotrzebowania na ciepło dla warszawskiej sieci ciepłowniczej na podstawie danych pomiarowych z systemu telemetrii oraz danych pogodowych.** Stwierdza się również, że stworzone modele prognostyczne, czyli kombinacja

sztucznej sieci neuronowej z autoregresyjnym wejściem wykorzystywanej dla pierwszych 24. godzin horyzontu predykcji oraz sztucznej sieci neuronowej bez autoregresyjnego wejścia wykorzystywanej dla kolejnych godzin horyzontu predykcji wykazują zadowalające wyniki dokładności predykcji, które są porównywalne lub lepsze z wynikami prezentowanymi w światowej literaturze naukowej i potwierdzają drugą tezę poboczną pracy. Wysoka jakość prezentowanych modeli prognostycznych ma szczególnie istotne znaczenie w kontekście wykorzystywania wyników prognozy w algorytmach optymalizujących pracę sieci ciepłowniczej, gdyż pozwalają na precyzyjne prognozowanie potrzeb odbiorców ciepła, a tym samym możliwe jest optymalne definiowanie pracy wytwórców ciepła tak, aby zaspokoić potrzeby konsumentów ciepła przy minimalnych stratach operacyjnych sieci ciepłowniczej.

Wartością dodaną prezentowaną w niniejszej pracy są prognozy zapotrzebowania na ciepło dla obszarów warszawskiej sieci ciepłowniczej, które stanowią dodatkową informację dla operatorów sieci. Modele prognostyczne dla obszarów sieci stworzone zostały na podstawie regresji grzbietowej, natomiast jakość generowanych prognoz dla większości obszarów jest porównywalna z jakościami prognoz wyznaczanych dla całej sieci, a na tej podstawie uważa się, że potwierdzono pierwszą tezę poboczną pracy.

Wysoka jakość generowanych prognoz pozwoliła na wdrożenie systemu prognozowania w Systemie Wsparcia Decyzji (SWD). Generowane prognozy stanowią zmienną wejściową algorytmu optymalizacji będącego podstawą funkcjonowania SWD. W rozdziale siódmym przedstawiono analizę danych z wdrożenia systemu oraz dwuletniej pracy, która potwierdza skuteczność prezentowanego rozwiązania zarówno w obszarze jakości generowanych prognoz, jak i w obszarze skuteczności wyznaczonych przedziałów ufności prognoz.

Dodatek A

Wkład osób trzecich w rozwój systemu prognostycznego

W załączniku przedstawiono wkład osób trzecich w rozwój systemu prognostycznego:

- Mgr inż. Artur Bielecki wspierał autorkę systemu podczas tworzenia finalnego modelu prognozującego zapotrzebowanie na ciepło dla całej sieci oraz przy implementacji wspomnianego modelu w systemi SWD. Pan mgr inż. Artur Bielecki wykonał trenowanie oraz optymalizację wartości hiperparametrów sztucznej sieci neuronowej wykorzystywanej w module prognostycznym.
- Mgr inż. Michał Guzek oraz Mgr inż. Jakub Białek stanowili wsparcie w obszarze wykorzystania prognoz generowanych przez system SPROG w algorytmie optymalizującym pracę warszawskiej sieci ciepłowniczej.
- Mgr inż. Rafał Brzozowski oraz mgr inż. Rafał Serafin jako pracownicy Veolia Energia Warszawa S.A. stanowili pomoc w obszarze zbierania danych pomiarowych z warszawskiej sieci ciepłowniczej. Ponadto mgr inż. Rafał Brzozowski oraz mgr inż. Rafał Serafin wspierali autorkę systemu wiedzą z zakresu funkcjonowania warszawskiej sieci ciepłowniczej.
- Dr inż. Konrad Wojdan jako dyrektor działu badawczo-rozwojowego w firmie Transition Technologies Science S.A. koordynował proces powstawania systemu SPROG oraz służył merytorycznym wsparciem w obszarze budowania oraz implementacji systemów opartych o techniki uczenia maszynowego.

Bibliografia

[1] *https://energiadlawarszawy.pl/strefa-miejska/jak-powstaje-cieplo/mapa-sieci-cieplowniczej/ [dostęp 09.11.2020]*.

[2] *http://www.ogrzewamyinteligentnie.pl/pl/33,warszawska-siec-cieplownicza.html [dostęp 09.11.2020]*.

[3] Dynamic time warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer Berlin Heidelberg, 2007.

[4] E. Abel. Low-energy buildings. *Energy and Buildings*, 21(3):169–174, jan 1994.

[5] A. C. Alice Zheng. *Feature Engineering for Machine Learning Models*. O'Reilly UK Ltd., 2018.

[6] K. T. Aminu, D. McGlinchey, and A. Cowell. Acoustic signal processing with robust machine learning algorithm for improved monitoring of particulate solid materials in a gas flowline. *Flow Measurement and Instrumentation*, 65:33–44, mar 2019.

[7] D. Aparicio and M. I. Bertolotto. Forecasting inflation with online prices. *International Journal of Forecasting*, 36(2):232–247, apr 2020.

[8] K. Baltputnis, R. Petrichenko, and A. Sauhats. ANN-based city heat demand forecast. In *2017 IEEE Manchester PowerTech*. IEEE, jun 2017.

[9] K. Baltputnis, R. Petrichenko, and D. Sobolevsky. Heating demand forecasting with multiple regression: Model setup and case study. In *2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*. IEEE, nov 2018.

[10] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *JMLR.org*, 2012.

[11] A. Boukerche and J. Wang. Machine learning-based traffic prediction models for intelligent transportation systems. *Computer Networks*, 181:107530, nov 2020.

[12] L. Breiman. *Machine Learning*, 45(1):5–32, 2001.

[13] H. Brink, M. Fetherolf, J. W. Richards, H. Brink, and J. Richards. *Real-World Machine Learning*. Manning, 2016.

[14] J. W. BULCOCK and W. F. LEE. Normalization ridge regression in practice. *Sociological Methods & Research*, 11(3):259–303, feb 1983.

[15] J. Cao, A. Kummert, Z. Lin, and J. Velten. Recent advances in machine learning for signal analysis and processing. *Journal of the Franklin Institute*, 355(4):1513–1516, mar 2018.

[16] N. Channouf, P. L'Ecuyer, A. Ingolfsson, and A. N. Avramidis. The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta. *Health Care Management Science*, 10(1):25–45, nov 2006.

[17] E. Chong, C. Han, and F. C. Park. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205, oct 2017.

[18] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, mar 2020.

[19] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini. The COVID-19 pandemic. *Critical Reviews in Clinical Laboratory Sciences*, 57(6):365–388, jul 2020.

[20] R. B. D'AGOSTINO. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971.

[21] M. Dahl, A. Brun, and G. B. Andresen. Using ensemble weather predictions in district heating operation and load forecasting. *Applied Energy*, 193:455–465, may 2017.

[22] M. Dahl, A. Brun, O. Kirsebom, and G. Andresen. Improving short-term heat load forecasts with calendar and holiday data. *Energies*, 11(7):1678, jun 2018.

[23] D. der Naturwissenschaften. Permutation distribution clustering and structural equation model trees. 2011.

[24] E. Dotzauer. Simple model for prediction of loads in district-heating systems. *Applied Energy*, 73(3-4):277–284, nov 2002.

[25] T. Fang and R. Lahdelma. Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system. *Applied Energy*, 179:544–552, oct 2016.

[26] H. Gadd and S. Werner. Daily heat load variations in swedish district heating systems. *Applied Energy*, 106:47–55, jun 2013.

[27] B. Y. Glorot X. Understanding the difficulty of training deepfeedforward neural networks. i. *ISTATS,*, 9:249–256, 2010.

[28] M. Gruber. *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. CRC Press, 2017.

[29] I. Guyon and A. Elisseeff. An introduction of variable and feature selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, 3:1157–1182, Jan. 2003.

[30] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, jun 2000.

[31] A. Heller. Heat-load modelling for large systems. *Applied Energy*, 72(1):371–387, may 2002.

[32] A. E. Hoerl and R. W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[33] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny. Application of big data and machine learning in smart grid, and associated security concerns: A review. *IEEE Access*, 7:13960–13988, 2019.

[34] W. Huang, X. Cao, and S. Zhong. Network-based comparison of temporal gene expression patterns. *Bioinformatics*, 26(23):2944–2951, sep 2010.

[35] S. Idowu, S. Saguna, C. Ahlund, and O. Schelen. Forecasting heat load for smart district heating systems: A machine learning approach. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, nov 2014.

[36] T. Jiang, J. L. Gradus, and A. J. Rosellini. Supervised machine learning: A brief primer. *Behavior Therapy*, 51(5):675–687, sep 2020.

[37] C. Johansson, M. Bergkvist, D. Geysen, O. D. Somer, N. Lavesson, and D. Vanhoudt. Operational demand forecasting in district heating systems using ensembles of online machine learning algorithms. *Energy Procedia*, 116:208–216, jun 2017.

[38] K. Kato, M. Sakawa, K. Ishimaru, S. Ushiro, and T. Shibano. Heat load prediction through recurrent neural network in district heating and cooling systems. In *2008 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, oct 2008.

[39] A. I. Khan and S. Al-Habsi. Machine learning in computer vision. *Procedia Computer Science*, 167:1444–1451, 2020.

[40] O. Khan, J. H. Badhiwala, G. Grasso, and M. G. Fehlings. Use of machine learning and artificial intelligence to drive personalized medicine approaches for spine care. *World Neurosurgery*, 140:512–518, aug 2020.

[41] A. Kosonen and A. Keskisaari. Zero-energy log house – future concept for an energy efficient building in the nordic conditions. *Energy and Buildings*, 228:110449, dec 2020.

[42] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer New York, 2013.

[43] T. Kurek, A. Bielecki, K. Świrski, K. Wojdan, M. Guzek, J. Białek, R. Brzozowski, and R. Serafin. Heat demand forecasting algorithm for a warsaw district heating network. *Energy*, 217:119347, feb 2021.

[44] T. Kurek, K. Wojdan, D. Nabagło, and K. Świrski. Detection of malfunctions and abnormal working conditions of a coal mill. In *Thermal Power Plants - New Trends and Recent Developments*. InTech, may 2018.

[45] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, jan 2016.

[46] H. Lund, N. Duic, P. A. Østergaard, and B. V. Mathiesen. Future district heating systems and technologies: On the role of smart energy systems and 4th generation district heating. *Energy*, 165:614–619, dec 2018.

[47] X. Luo, J. Lu, and J. Ge. Green and energy-saving reform technology research in traditional houses – taking luo's house in hangzhou as an example. *Journal of Asian Architecture and Building Engineering*, pages 1–14, jul 2020.

[48] H. M, G. E.A., V. K. Menon, and S. K.P. NSE stock market prediction using deep-learning models. *Procedia Computer Science*, 132:1351–1362, 2018.

[49] V. Marinakis. Big data for energy management and energy-efficient buildings. *Energies*, 13(7):1555, mar 2020.

[50] B. Möller, E. Wiechers, U. Persson, L. Grundahl, R. S. Lund, and B. V. Mathiesen. Heat roadmap europe: Towards EU-wide, local heat supply strategies. *Energy*, 177:554–564, jun 2019.

[51] T. Nigitz and M. Gölles. A generally applicable, simple and adaptive forecasting method for the short-term heat load of consumers. *Applied Energy*, 241:73–81, may 2019.

[52] J. Noorbakhsh, H. Chandok, R. K. M. Karuturi, and J. George. Machine learning in biology and medicine. *Advances in Molecular Pathology*, 2(1):143–152, nov 2019.

[53] J. Owen, B. Eccles, B. Choo, and M. Woodings. The application of auto–regressive time series modelling for the time–frequency analysis of civil engineering structures. *Engineering Structures*, 23(5):521–536, may 2001.

[54] D. Park, M. El-Sharkawi, R. Marks, L. Atlas, and M. Damborg. Electric load forecasting using an artificial neural network. *IEEE Transactions on Power Systems*, 6(2):442–449, may 1991.

[55] N. Perez-Mora, V. Martinez-Moll, and V. Canals. DHC load management using demand forecast. *Energy Procedia*, 91:557–566, jun 2016.

[56] U. Persson, E. Wiechers, B. Möller, and S. Werner. Heat roadmap europe: Heat distribution costs. *Energy*, 176:604–622, jun 2019.

[57] R. A. D. Peter J. Brockwell. *Introduction to Time Series and Forecasting*. Springer; 2nd edition, 2010.

[58] R. Petrichenko, K. Baltputnis, A. Sauhats, and D. Sobolevsky. District heating demand short-term forecasting. In *2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*. IEEE, jun 2017.

[59] P. Potočnik, E. Strmčnik, and E. Govekar. Linear and neural network-based models for short-term heat load forecasting. *Strojniški vestnik – Journal of Mechanical Engineering*, 61(9):543–550, sep 2015.

[60] S. RASCHKA. *Python Machine Learning*. Packt Publishing Limited, 2015.

[61] D. Raspopov and P. Belousov. Development of methods and algorithms for identification of a type of electric energy consumers using artificial intelligence and machine learning models for smart grid systems. *Procedia Computer Science*, 169:597–605, 2020.

[62] K. H. Refat and R. N. Sajjad. Prospect of achieving net-zero energy building with semi-transparent photovoltaics: A device to system level perspective. *Applied Energy*, 279:115790, dec 2020.

[63] B. Y. Reis and K. D. Mandl. Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, 3(1), jan 2003.

[64] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. A. Brown, and R. M. Pendyala. Machine learning approaches for estimating commercial building energy consumption. *Applied Energy*, 208:889–904, dec 2017.

[65] A. Sandberg, F. Wallin, H. Li, and M. Azaza. An analyze of long-term hourly district heat demand forecasting of a commercial building using neural networks. *Energy Procedia*, 105:3784 – 3790, 2017. 8th International Conference on Applied Energy, ICAE2016, 8-11 October 2016, Beijing, China.

[66] L. B. Sheremetov, A. González-Sánchez, I. López-Yáñez, and A. V. Ponomarev. Time series forecasting: Applications to the upstream oil and gas supply chain. *IFAC Proceedings Volumes*, 46(9):957–962, 2013.

[67] J. Smith. *Reactive Machine Learning Systems*. Manning, 2018.

[68] M. Sobczyk. *Prognozowanie. Teoria. Przykłady. Zadania*. Placet, 2008.

[69] J. Soni, U. Ansari, D. Sharma, and S. Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, mar 2011.

[70] G. Suryanarayana, J. Lago, D. Geysen, P. Aleksiejuk, and C. Johansson. Thermal load forecasting in district heating networks using deep learning and advanced feature selection methods. *Energy*, 157:141–149, aug 2018.

[71] K. Szafranek. Bagged neural networks for forecasting polish (low) inflation. *International Journal of Forecasting*, 35(3):1042–1059, jul 2019.

[72] L. J. Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, oct 2000.

[73] H. Tommerup, J. Rose, and S. Svendsen. Energy-efficient houses built according to the energy performance requirements introduced in denmark in 2006. *Energy and Buildings*, 39(10):1123–1130, oct 2007.

[74] H. Wiklund. Short term forecasting on the heat load in a dh-system. *District heating international*, 20:286–294, 1991.

[75] K. Wojdyga. Predicting heat demand for a district heating systems. *International Journal of Energy and Power Engineering*, 3(5):237, 2014.

[76] G. Xue, Y. Pan, T. Lin, J. Song, C. Qi, and Z. Wang. District heating load prediction algorithm based on feature fusion LSTM model. *Energies*, 12(11):2122, jun 2019.

[77] L. Yin, Q. Gao, L. Zhao, B. Zhang, T. Wang, S. Li, and H. Liu. A review of machine learning for new generation smart dispatch in power systems. *Engineering Applications of Artificial Intelligence*, 88:103372, feb 2020.

[78] A. Zheng. *Evaluating Machine Learning Models*. O'Reilly Media, 2015.

[79] H. Zhou, B. Lin, J. Qi, L. Zheng, and Z. Zhang. Analysis of correlation between actual heating energy consumption and building physics, heating system, and room position using data mining approach. *Energy and Buildings*, 166:73–82, may 2018.

[80] E. Zvingilaite. Modelling energy savings in the danish building sector combined with internalisation of health related externalities in a heat and power system optimisation model. *Energy Policy*, 55:57–72, apr 2013.