http://dx.doi.org/10.18778/7525-926-1

Modelowanie niepewności krótkoterminowego popytu na energię elektryczną z wykorzystaniem sieci neuronowych i neuronowo-rozmytych

WYDAWNICTWA
UNIWERSYTETU
ŁÓDZKIEGO

WITOLD BARTKIEWICZ

Modelowanie niepewności krótkoterminowego popytu na energię elektryczną z wykorzystaniem sieci neuronowych i neuronowo-rozmytych

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

ŁÓDŹ 2013

Witold Bartkiewicz – Katedra Informatyki, Wydział Zarządzania
Uniwersytet Łódzki, 90-237 Łódź, ul. Matejki nr 22/26
e-mail: wbartkiewicz@wzmail.uni.lodz.pl

RECENZENT
Gabriela Idzikowska

OPRACOWANIE REDAKCYJNE
Urszula Dzieciątkowska

SKŁAD I ŁAMANIE
Leonora Wojciechowska

PROJEKT OKŁADKI
Barbara Grzejszczak

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.06235.13.0.H

ISBN (wersja drukowana) 978-83-7525-926-1
ISBN (ebook) 978-83-7969-154-8

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63, faks (42) 665 58 62

Spis treści

Wstęp

Rozwój mechanizmów konkurencyjnego rynku energii elektrycznej w Polsce stawia przed przedsiębiorstwami elektroenergetycznymi poważne wyzwania biznesowe. Skuteczne funkcjonowanie w tej dziedzinie gospodarki wymaga uwzględnienia wielu czynników ryzyka, wynikających w dużej mierze z niepewności rynkowej, szczególnie dotkliwej w przypadku obrotu takimi towarami jak energia elektryczna. W związku z tym w działalności rynkowej przedsiębiorstw tego sektora właściwe zarządzanie ryzykiem nabiera kluczowego znaczenia, decydującego wręcz o wyniku finansowym.

Podstawowym elementem każdego systemu zarządzania ryzykiem musi być identyfikacja oraz właściwa ocena generujących je czynników niepewności. Z ryzykiem mamy bowiem do czynienia wyłącznie w sytuacji, w której podejmowane decyzje mogą dać różne skutki, w tym również niepożądane, wiążące się ze stratami lub gorszymi efektami realizowanych działań. Istotne jest przy tym, że zazwyczaj skutki te potrafimy określić jedynie nieprecyzyjnie, np. z dokładnością do prawdopodobieństwa ich występowania. Ryzyko jest więc nieodłącznie związane z niepewnością. Jeżeli efekty decyzji są pewne, ryzyko nie występuje. Możliwość modelowania niepewności, określania prawdopodobieństw niepewnych zjawisk, które wpływają na skutki podejmowanych działań, pozwalają zatem włączać ryzyko w proces podejmowania decyzji czy też zabezpieczać się przed jego efektami.

Cechy charakterystyczne energii elektrycznej jako towaru, takie jak brak praktycznych możliwości jej magazynowania na poważniejszą skalę, konieczność nieustannego równoważenia wytwarzania i odbioru energii, powodują występowanie na rynkach energii szybkich zmian cen oraz zapotrzebowania. Niepewność popytowa stanowi więc jeden z podstawowych czynników wpływających na powstawanie ryzyka działania przedsiębiorstwa energetycznego. Rynek energii zdominowany jest w znacznej mierze przez kontrakty terminowe o średnich i dłuższych okresach realizacji, które zastąpiły w dużej części dawne kontrakty dwustronne. Rozwój takich segmentów, jak rynek giełdowy czy też rynek bilansujący oraz wzrost konkurencji w obrębie sektora wywołują jednak pewne przesunięcie działań handlowych w kierunku transakcji o horyzontach krótkoterminowych. Wyraźnie widoczny jest zwłaszcza rozwój rynków dnia następnego. Oferują one przedsiębiorstwu większą elastyczność działania,

skutkują jednak m.in. znaczniejszą ekspozycją na ryzyko wynikające z niepewności popytowej.

Dążenie do redukcji tej niepewności jest jednym z głównych powodów usilnego poszukiwania jak najdokładniejszych metod krótkoterminowego prognozowania zapotrzebowania na energię elektryczną. Zmniejszenie błędu w oszacowaniu popytu nawet o ułamek procenta przekłada się bowiem na wymierną kwotę w wynikach finansowych przedsiębiorstwa energetycznego. Zaawansowane metody modelowania danych stanowią obecnie standardowy element systemów prognozowania zapotrzebowania odbiorców na rynku energii. Wśród nich do powszechnie wykorzystywanych należą metody oparte na sieciach neuronowych i neuronowo-rozmytych.

Nieustanna pogoń za poprawianiem dokładności prognozy powoduje koncentrację wysiłków badawczych na prognozach punktowych, na uzyskiwaniu coraz lepszego oszacowania wartości oczekiwanej procesu zapotrzebowania na energię. Jest to, oczywiście, w znacznej mierze postępowanie uzasadnione. Pomija się jednak przy tym często fakt, że z punktu widzenia procesu decyzyjnego, znajomość wyłącznie wartości oczekiwanej zapotrzebowania na energię w wielu przypadkach może być niewystarczająca. Najlepsza bowiem nawet prognoza stanowi jedynie oszacowanie, obarczone niepewnością. Kwestia modelowania tej niepewności, określenia rozkładu prawdopodobieństwa prognozy dla danego wzorca wejściowego, jest często zaniedbywana, a przecież stanowi ona podstawowy element oszacowania ryzyka działań i decyzji opierających się na sporządzonej prognozie.

W niniejszej pracy próbujemy uzupełnić tę lukę, badając zaprezentowaną problematykę. Podstawowa teza, jaką stawiamy, brzmi:

Współczesne realia funkcjonowania rynku energii elektrycznej spowodowały, że zaawansowane metody modelowania, takie jak sieci neuronowe i neuronowo-rozmyte, weszły do standardowego zestawu narzędzi stosowanych w procesie krótkoterminowego prognozowania zapotrzebowania na energię. Obok prognoz wartości oczekiwanej popytu odbiorców, ocena ryzyka decyzji opierających się na tych prognozach i włączenie tego ryzyka w proces podejmowania decyzji wymagają zastosowania metod modelowania niepewności prognozowanego zapotrzebowania, w postaci oszacowania rozkładu warunkowego prognozy dla danego wzorca wejściowego.

Pamiętajmy również, że sieci neuronowe i neuronowo-rozmyte należą do indukcyjnych modeli nieliniowych o bogatej strukturze aproksymacyjnej. Dla tego rodzaju narzędzi nie ma prostych metod szacowania warunkowej wariancji wyjściowej modelu, przy danym wejściu. Stosowane są tu podejścia empiryczne lub podejścia analityczne o przybliżonym charakterze, wymagającym również weryfikacji empirycznej w każdym konkretnym przypadku zastosowań. Istotnym problemem może być więc stosowalność określonych algorytmów

w zadaniu krótkoterminowego prognozowania zapotrzebowania na energię elektryczną.

Przeprowadzone badania pozwalają na postawienie w prezentowanej pracy dodatkowej pomocniczej hipotezy badawczej dotyczącej tego zagadnienia:

Badane metody szacowania warunkowej wariancji wyjściowej, dla danego wzorca wejściowego, stanowią odpowiednie narzędzia, które mogą być rozważane przy modelowaniu niepewności krótkoterminowych prognoz zapotrzebowania na energię elektryczną, w przypadku zastosowania analizowanych w pracy neuronowych i neuronowo-rozmytych metod prognozowania.

Presja na poprawę dokładności krótkoterminowych prognoz zapotrzebowania na energię elektryczną spowodowała, że niemal od razu po wprowadzeniu, w drugiej połowie lat osiemdziesiątych ubiegłego wieku, algorytmu wstecznej propagacji błędu i opracowaniu modelu wielowarstwowej sieci perceptronowej (MLP), sieć ta znalazła swoje zastosowanie również i w tych zagadnieniach. Przełom lat osiemdziesiątych i dziewięćdziesiątych XX w. to prawdziwa eksplozja zastosowań sztucznych sieci neuronowych w prognozowaniu w elektroenergetyce. Wymienić można tutaj choćby takie prace, jak: Peng, Hubele, Karady 1990; Dillon, Sestito, Leung 1991; El-Sharkawi, Oh, Marks, Damborg, Brace 1991; Lee, Cha, Ku 1991; Srinivasan, Liew, Chen 1991; Park, El-Sharkawi, Marks, Atlas, Damborg 1991. Te pierwsze zastosowania związane były przede wszystkim z wykorzystaniem do krótkoterminowej predykcji zapotrzebowania na energię nieliniowych właściwości sieci MLP, pozwalających na lepsze modelowanie zależności procesu zapotrzebowania od czynników pogodowych. Stosunkowo proste modele neuronowe działały z co najmniej porównywalną, a zazwyczaj lepszą, dokładnością niż złożone i wyrafinowane komercyjne systemy prognostyczne.

Na połowę lat dziewięćdziesiątych ubiegłego wieku datuje się również pierwsze zastosowania sztucznych sieci neuronowych do prognozowania krótkoterminowego zapotrzebowania na energię w polskich systemach elektroenergetycznych. Wymienić można tu dla przykładu takie prace, jak: Heine, Malko, Mikołajczak, Skorupski 1994; Malko, Mikołajczak, Skorupski 1995; Nazarko, Jurczuk 1996, w tym również prace prowadzone i współprowadzone przez autora niniejszej rozprawy: Bardzki, Bartkiewicz 1995; Bardzki, Bartkiewicz, Zieliński 1995.

Niemal od początku szukano jednak dalszej poprawy dokładności prognoz i efektywności działania modelu na drodze korekt w algorytmie uczenia sieci neuronowej (np. Ho, Hsu, Yang 1992) czy też drobnych modyfikacji architektury modelu, takich jak, dla przykładu, uproszczenie struktury połączeń między węzłami (Chen, Yu, Moghaddamjo 1992).

Istotnym kierunkiem poszukiwań lepszych prognoz krótkoterminowego zapotrzebowania na energię elektryczną jest hybrydyzacja modelu poprzez

połączenie metod prognozowania o różnych właściwościach, tak by wyeliminować słabe strony każdej z nich. Koronnym przykładem systemów hybrydowych są oczywiście sieci neuronowo-rozmyte; tym zagadnieniem zajmiemy się nieco dalej. Możliwości hybrydyzacji sieci neuronowych z innymi technikami prognostycznymi jest jednak wiele. Do najważniejszych kierunków zaliczyć tutaj możemy: podział modelu prognostycznego na grupę modeli o charakterze lokalnym, działających w obrębie bardziej jednorodnych segmentów danych; łączenie sieci neuronowych z dedukcyjnymi technikami analizy danych, co pozwala wykorzystać w pewnej części właściwości klasycznego wnioskowania statystycznego; łączenie sieci neuronowych z systemami ekspertowymi czy też ogólniej – systemami opartymi na wiedzy.

Pierwsze prace nad tworzeniem modeli lokalnych związane były z wykorzystaniem sieci Kohonena, czy też innych technik nienadzorowanego grupowania danych, do podziału całego zbioru obserwacji historycznych na bardziej jednorodne segmenty. Dla przykładu, Hsu, Yang (1991a, b) sieć Kohonena wykorzystują do wstępnej klasyfikacji dni o podobnych wzorcach dobowego zapotrzebowania na energię elektryczną. Profil danej grupy stanowi dalej wzorzec wejściowy dla prognozy MLP. Podobne podejście prezentowane jest również w innych pracach: Erkmen, Ozsokmen 1993; Bardzki, Bartkiewicz 1995. W bardziej rozwiniętej postaci mamy w tym podejściu do czynienia z grupą lokalnych predyktorów, których działanie scalane jest w jedną wspólną prognozę (np. Drezga, Rahman 1999).

Hybrydyzacja sieci neuronowych z klasycznymi technikami analizy danych związana jest z podziałem modelu prognostycznego na część indukcyjną, w której wykorzystuje się sieć neuronową, oraz część dedukcyjną, co do której istniejąca wiedza umożliwia poczynienie pewnych założeń pozwalających na stosowanie klasycznych, bardziej efektywnych i prostszych w implementacji modeli wnioskowania statystycznego. Przykładem może być tu rozdział modelu na część autoregresyjną oraz zależną od danych meteorologicznych (Peng, Hubele, Karady 1992; Taylor 2012) czy też wykorzystanie metod analizy szeregów czasowych do badania reszt modelu (Wang, Xia, Kang 2011). Innym przykładem jest z kolei łączenie sieci neuronowych z metodami analizy częstotliwościowej i filtrowania, takimi jak np. analiza falkowa (np. Huang, Yang 2001; Pandey, Singh, Sinha 2010).

Hybrydy sieci neuronowych z klasycznymi systemami opartymi na wiedzy związane są przede wszystkim z próbą uwzględnienia ingerencji w model predykcji doświadczonego operatora dokonującego korekt prognozy zapotrzebowania na energię na podstawie jakościowych czynników pogodowych czy też informacji o niespodziewanych zdarzeniach wpływających na poziom zapotrzebowania. Za przykład posłużyć tu może połączenie sieci neuronowej z systemem ekspertowym (Rahman, Hazim 1993) albo drzewem decyzyjnym (Rahman, Drezga, Rajagopalan 1993). Innym rozwiązaniem, które można zaliczyć do tej

grupy, jest tworzenie modeli dla różnych typów dni, z wykorzystaniem reguł heurystycznych odpowiedniego doboru wejść (Bardzki, Bartkiewicz, Gontar, Zieliński 1998; Bartkiewicz, Gontar, Zieliński, Bardzki 2000a, b).

W początkowym okresie zastosowań sieci neuronowych w zadaniach krótkoterminowego prognozowania zapotrzebowania na energię elektryczną stosowano głównie wielowarstwowe sieci perceptronowe (MLP). Inne architektury wykorzystywane były epizodycznie bądź miały charakter pomocniczy w systemach hybrydowych. Stopniowo jednak, obok modelu MLP, duże znaczenie w dyskutowanych zagadnieniach zaczęły zyskiwać sieci z funkcjami o bazie radialnej (RBF) (np. Ranaweera, Hubele, Papalexopoulos 1995; Gontar, Hatziargyriou 2001; Gontar, Sideratos, Hatziargyriou 2004; Zhang, Zhou, Sun, Lei, Liu, Song 2008). W niniejszej pracy nie omawiamy, co prawda, odrębnie sieci RBF, jednakże jeden z analizowanych modeli neuronowo-rozmytych (FBF) ma strukturę bardzo do nich zbliżoną i można wykazać funkcjonalną równoważność tych dwu podejść (Jang, Sun 1993).

Najbardziej produktywną chyba metodą hybrydyzacji sieci neuronowych jest ich łączenie z systemami z logiką rozmytą. Architektury neuronowo-rozmyte zaczęły pojawiać się w zadaniach krótkoterminowej prognozy zapotrzebowania na energię niemal równocześnie z modelami neuronowymi. Początkowo miały one charakter luźnych powiązań, na ogół o charakterze sekwencyjnym. Podstawowa prognoza sporządzana była przez sieć neuronową, a następnie korygowana przez system z logiką rozmytą, zazwyczaj na bazie informacji pogodowej. Możliwa była odwrotna kolejność, tj. model rozmyty analizował dane pogodowe, tworząc wejścia dla sieci neuronowej. Przykładowe rozwiązania tego rodzaju prezentowane są w: Lambert-Torres, Traore, Mandolesi, Mukhedkar 1991; Kim, Park, Hwang, Kim 1995; Dash, Dash, Rahman 1993.

Dosyć szybko jednak w interesujących nas zagadnieniach krótkoterminowej prognozy zapotrzebowania na energię elektryczną zaczęto stosować modele hybrydowe o ścisłej integracji, nazywane sieciami neuronowo-rozmytymi. Pojęcie to ma dosyć szeroki zakres i obejmuje całą gamę rozwiązań hybrydowych. W praktyce największe znaczenie, przynajmniej w zadaniach prognozy zapotrzebowania na energię, mają sieci neuronowo-rozmyte w formie adaptacyjnych systemów z logiką rozmytą. Podejście to polega na takim ujęciu mechanizmu działania systemu rozmytego, aby parametry definiujące wykorzystywane w nim zbiory rozmyte mogły być uczone na podstawie danych.

Już w połowie lat dziewięćdziesiątych ubiegłego wieku zaczęto stosować systemy prognozowania krótkoterminowego zapotrzebowania na energię elektryczną, w których wykorzystuje się sieci neuronowo-rozmyte realizujące tzw. uproszczone wnioskowanie rozmyte (Mori, Kobayashi 1995; Bakirtzis, Theocharis, Kiartzis, Satsios 1995; Bartkiewicz 1998b, c; Bartkiewicz, Zieliński 1998; Mastorocostas, Theocharis, Bakirtzis 1999; Bartkiewicz, Butkevych i inni

2001). Tego rodzaju modele nazywa się często sieciami z funkcjami o bazie rozmytej (FBF).

Wkrótce w zadaniach krótkoterminowej prognozy zapotrzebowania na energię zaczęto również stosować sieci neuronowo-rozmyte realizujące wnioskowanie rozmyte typu Takagi–Sugeno. W modelach tego rodzaju w następnikach reguł rozmytych występują funkcje zmiennych wejściowych (Wu, Lu 1999; Bartkiewicz 2000d; Bartkiewicz, Butkevych i inni 2001; Zhang, Zhou, Sun, Lei, Liu, Song 2008; Hanmandlu, Chauhan 2011). Sieci, w których wykorzystuje się to podejście, często określane są również jako ANFIS.

Jak widzimy, sieci neuronowe i neuronowo-rozmyte stanowią dobrze już ugruntowane metody prognozowania krótkoterminowego zapotrzebowania na energię elektryczną. Wysoka jakość ich działania i skuteczność spowodowały, że większość komercyjnych systemów prognostycznych w interesującej nas dziedzinie korzysta z któregoś z tych rozwiązań jako z podstawowej techniki modelowania danych. Dlatego tematem niniejszej pracy nie jest zastosowanie sieci neuronowych i neuronowo-rozmytych w prognozowaniu zapotrzebowania. Nasz cel polega na przebadaniu problemu modelowania niepewności tych popularnych metod prognostycznych w rozważanych zagadnieniach.

Dotychczasowe badania w dziedzinie modelowania niepewności prognoz krótkoterminowego zapotrzebowania na energię elektryczną koncentrują się głównie na konstrukcji przedziałów prognozy dla sieci neuronowych. Można tu wymienić (obok badań autora) takie prace, jak: Ding 1999; da Silva, Moulin 2000; Khosravi, Nahavandi, Creighton 2010; Petiau 2009. Mają one na ogół wycinkowy charakter, brak jest szerszych opracowań porównawczych. Metody modelowania niepewności wyjścia nieliniowych modeli prognostycznych, takich jak sieci neuronowe i neuronowo-rozmyte, mają przybliżony lub empiryczny charakter i wymagają badań na różnorodnych problemach i zbiorach danych z danej dziedziny. Stąd właśnie wynika znaczenie prezentowanej pracy dla zagadnień krótkoterminowej prognozy zapotrzebowania na energię elektryczną. Rozprawa ta wypełnia poważną lukę – przedstawiono w niej szeroki zakres badań porównawczych dla różnych modeli prognostycznych i problemów z tej dziedziny.

By dowieść postawionych w pracy tez, jej zawartość podzielona została na cztery rozdziały.

W rozdziale pierwszym scharakteryzujemy właściwości energii elektrycznej jako towaru – zbadamy jej znaczenie dla funkcjonowania społeczeństwa oraz wskażemy na elementy wpływające na ryzyko popytowe. W dalszej części rozdziału przeanalizujemy konstrukcję rynku energii, omawiając mechanizm działania jego poszczególnych segmentów oraz najważniejsze procedury decyzyjne związane z funkcjonowaniem na nim przedsiębiorstw elektroenergetycznych.

W rozdziale drugim zaprezentujemy szeroki zestaw modeli neuronowych i neuronowo-rozmytych w zastosowaniu do krótkoterminowych prognoz zapotrzebowania na energię elektryczną. W rozdziale tym przedstawione zostaną wyniki badań wskazujących na wysoką dokładność analizowanych metod w porównaniu z innymi standardowymi technikami prognostycznymi.

Rozdział trzeci poświęcimy przedstawieniu najważniejszych metod określania warunkowego rozkładu prawdopodobieństwa prognozy, przy danym wzorcu wejściowym, dla analizowanych modeli neuronowych i neuronowo-rozmytych. Zaprezentujemy w nim również wyniki badań wykorzystania analizowanych metod w zagadnieniach krótkoterminowej prognozy zapotrzebowania na energię elektryczną. Uzyskane rozkłady prawdopodobieństwa zweryfikowano empirycznie, by ocenić ich zgodność z obserwowanymi danymi. Jako problem porównawczy wykorzystamy zadanie szacowania przedziałów prognozy na określonych poziomach prawdopodobieństwa.

Ostatni rozdział, czwarty, zawiera przegląd najważniejszych typów problemów decyzyjnych oraz sposobu wykorzystania w nich modeli niepewności zapotrzebowania na energię elektryczną. W rozdziale tym wskażemy na niedostateczny charakter prognoz punktowych, wartości oczekiwanej zapotrzebowania oraz na znaczenie wykorzystania informacji o niepewności prognozy pochodzącej z całości jej rozkładu prawdopodobieństwa.

Na koniec poczyńmy jeszcze dwie istotne uwagi dotyczące zakresu niniejszej pracy. Po pierwsze, rynek energii elektrycznej jest systemem, w którym silnie wiążą się ze sobą zagadnienia ekonomiczne oraz techniczne obrotu energią. W rozprawie tej ograniczamy się przede wszystkim do wykorzystania prognoz popytu w problemach ekonomicznych. Aspekty techniczne poruszane są jedynie w takim zakresie, aby pomóc zrozumieć Czytelnikowi pewne zagadnienia funkcjonowania rynku.

Po drugie, problemy prognozowania krótkoterminowego zapotrzebowania na energię elektryczną mają charakter uniwersalny, z punktu widzenia miejsca ich zastosowań. W rozważanych w prezentowanej pracy zagadnieniach decyzyjnych i wielu elementach charakterystyki rynku koncentrujemy się jednak raczej na punkcie widzenia przedsiębiorstw obrotu energią, czy też ogólnie – odbiorców kupujących energię na rynku hurtowym. Nie rozważamy specyficznych problemów związanych z wytwarzaniem energii, kosztami działania jednostek wytwórczych, ich charakterystykami działania itp.

ROZDZIAŁ 1

Rynek energii elektrycznej

Energia elektryczna, z powodu swych cech fizycznych oraz ze względu na znaczenie dla funkcjonowania współczesnego społeczeństwa, stanowi specyficzny towar, wymagający zastosowania specjalnych mechanizmów rynkowych, które zapewnią niezawodność fizycznej realizacji zawieranych umów handlowych. W bieżącym rozdziale analizujemy więc właściwości energii elektrycznej jako towaru, wykazując fundamentalny charakter niepewności popytowej jako źródła ryzyka operacji na rynku energii. Jak pokażemy dalej, nieodłączną cechą obrotu energią elektryczną jest, z jednej strony, konieczność jego ścisłego planowania i ustalania z góry wolumenów obrotu, w formie zawieranych kontraktów handlowych, z drugiej zaś, równie nieodłączna niemożność dokładnej realizacji zawartych umów spowodowana niepewnością zapotrzebowania. Znaczenie tego problemu jest na tyle poważne, że konieczność jego rozwiązania w zasadzie determinuje samą konstrukcję współczesnych rynków energii.

Przedstawiona w rozdziale analiza tych zagadnień pozwoli nam udowodnić elementy tezy pracy wskazujące na znaczenie prognozowania krótkoterminowego zapotrzebowania na energię jako mechanizmu redukcji tak istotnego czynnika niepewności popytowej. Przyjrzymy się również strukturom i mechanizmom obrotu na rynkach energii, co pozwoli nam określić i wskazać elementy ryzyka działań handlowych, wynikającego z niepewności prognoz zapotrzebowania, a zarazem uświadomi, wskazywaną w tezie pracy, konieczność modelowania tejże niepewności dla osiągnięcia poprawy jakości procesów decyzyjnych w tej dziedzinie.

1.1. Ogólna charakterystyka procesu handlu energią elektryczną

1.1.1. Energia – przeszłość, teraźniejszość i przyszłość

Energia stanowi jedno z najważniejszych dóbr decydujących o poziomie rozwoju cywilizacyjnego społeczeństwa. Pozwala ona na odciążenie ludzi od nużących prac fizycznych i w ten sposób nie tylko poprawia ich warunki życia, ale zwiększa także ilość czasu, który mogą oni przeznaczyć na swój rozwój na różnych płaszczyznach życia. W tabeli 1.1.1 prezentujemy poglądowe oszaco-

wania ilości energii wykorzystywanej w różnych fazach rozwoju cywilizacyjnego w podstawowych dziedzinach życia. Jak widzimy, energia wykorzystywana była przez ludzkość od zarania jej dziejów. Początkowo oczywiście była to energia niezbędna do pozyskania i przygotowania żywności, stopniowo jej wykorzystanie rozszerzyło się na gospodarstwa domowe (ogrzewanie, oświetlanie) oraz handel, a następnie na przemysł, rolnictwo, transport i inne dziedziny życia. Zużycie energii przez ludzkość rośnie wyraźnie wraz z rozwojem cywilizacyjnym. Dobro to wykorzystywane jest coraz szerzej, co z jednej strony stwarza nowe perspektywy rozwojowe, z drugiej jednak strony, uzależnia społeczeństwo od stałej, pewnej i taniej jego dostępności.

Za bezsporny fakt należy więc uznać, że energia była, jest i będzie jednym z najważniejszych dóbr niezbędnych do istnienia społeczeństwa. Trudno sobie wyobrazić funkcjonowanie niemal każdego aspektu dzisiejszej cywilizacji: nowoczesnej telekomunikacji i transportu, działalności biznesowej, produkcyjnej, usługowej lub handlowej, bez możliwości sprawnego i niezawodnego korzystania z zasobów energetycznych. Równie trudne do zaakceptowania byłoby życie codzienne ludzi, bez dostępu do bezpiecznej, czystej i taniej energii, energii do celów oświetleniowych, grzewczych, do zasilania różnorodnych urządzeń wykorzystywanych w gospodarstwach domowych. Jak uczą nas doświadczenia wynikające z wydarzeń nadzwyczajnych – które skutkują krótkotrwałym i ograniczonym terytorialnie brakiem zasilania w energię (zwłaszcza elektryczną) – wywołanych klęskami naturalnymi lub rozległymi awariami, brak dostępu do energii na większą skalę oznaczałby po prostu załamanie cywilizacyjne.

Tabela 1.1.1. Zapotrzebowanie na energię na różnych poziomach rozwoju cywilizacyjnego ludzkości

Epoka	Dzienne zużycie (tys. kcal)				
	Żywność	Gospodarstwa domowe i handel	Przemysł i rolnictwo	Transport	Ogółem
Pierwotna	2				2
Łowiecka	3	2			5
Pierwotna rolnicza	4	4	4		12
Zaawansowana rolnicza	6	12	7	1	26
Przemysłowa	7	32	24	14	77
Technologiczna	10	66	91	63	230

Źródło: na podstawie T. Witkowski, *Energia – możliwości naukowe i bariery technologiczne oraz społeczne*, „Czysta Energia" 2011, nr 5, za: E. Cook, *The Flow of Energy in an Industrial Society*, „Scientific American" 1971, no. 225(3), s. 135–144.

Oczywiście energia elektryczna pojawia się na scenie dziejów dopiero w późniejszym okresie. Początkowo podstawowe formy energii wykorzystywane przez człowieka to przede wszystkim energia cieplna, powstająca ze spalania drewna i innych substancji naturalnych, energia mechaniczna czy też energia wytwarzana przez udomowione zwierzęta. Doniesienia o wykorzystywaniu energii cieplnej pary wodnej pojawiają się już w starożytności, ale jej rzeczywiste zastosowanie na szeroką skalę datuje się na koniec XVIII w. i wiąże się z wynalezieniem przez Jamesa Watta maszyny parowej oraz z rewolucją przemysłową w Anglii. O praktycznym wykorzystaniu energii elektrycznej tak naprawdę możemy mówić odnośnie do początku XIX w., kiedy to w 1800 r. Alessandro Volta buduje pierwsze ogniwo galwaniczne.

W ciągu stu kilkudziesięciu lat energia, w tym również energia elektryczna, stała się jednym z podstawowych dóbr, od których zależy funkcjonowanie społeczeństwa. W tym okresie wraz z rozwojem gospodarczym i cywilizacyjnym świata w szybkim tempie wzrastało także zapotrzebowanie na energię. W obliczu rosnącego uzależnienia świata od wykorzystania zasobów energetycznych pojawiają się oczywiście niepokojące pytania na temat perspektyw dalszego rozwoju sytuacji, zwłaszcza w kontekście takich zagadnień, jak dostępność zasobów surowców energetycznych, kwestie zanieczyszczenia środowiska naturalnego czy emisji gazów cieplarnianych i ocieplenia klimatu.

Paradoksalnie w chwili obecnej wymienione czynniki powodują, że pożądaną perspektywą dalszego rozwoju współczesnej cywilizacji staje się raczej dążenie do redukcji zapotrzebowania na energię i znalezienie innych, „czystych", źródeł jej produkcji. Takie zagadnienia jak zwiększenie efektywności wykorzystania energii, nie tylko w gospodarce, ale również w życiu codziennym ludzi, czy energia ze źródeł odnawialnych, stanowią jedne z najważniejszych wyzwań stojących przed współczesną energetyką i ogólnie przed światową cywilizacją.

1.1.2. Cechy charakterystyczne energii elektrycznej jako towaru

Zasadniczy cel istnienia przedsiębiorstwa można określić jako osiąganie zysku poprzez zaopatrywanie klientów w określone produkty. Wśród podstawowych funkcji organizacji gospodarczych wymienić więc można funkcję produkcyjną, rozumianą jako użytkowanie różnego rodzaju materiałów, środków technicznych i usług w celu wytworzenia nowych produktów i usług wymaganych przez klienta (Durlik 1995). Ponieważ potrzeby konsumentów znajdują się w centrum procesu produkcji, zdolność ich trafnego przewidywania i zaspokajania warunkuje użyteczność istnienia organizacji.

Wyjątku od tej reguły nie stanowią przedsiębiorstwa elektroenergetyczne. I tak w ich przypadku podstawowe cele działania wiążą się z zapewnieniem

odbiorcom dostaw energii elektrycznej o odpowiednich parametrach, w sposób ciągły oraz bezpieczny, w powiązaniu oczywiście z osiąganiem jak największych zysków z tejże działalności. Należy jednak wziąć pod uwagę pewne istotne specyficzne cechy funkcjonowania systemu elektroenergetycznego (a co za tym idzie, rynku energii elektrycznej), które powodują, że problem trafnego określenia zapotrzebowania odbiorców na energię nabiera podstawowego znaczenia dla efektywnego i poprawnego działania całości systemu oraz wpływa na konstrukcję rozwiązań rynkowych w tej dziedzinie.

Wysoka kosztochłonność inwestycji w infrastrukturę elektroenergetyczną w większości przypadków wymusza, rzecz jasna, rezygnację ze stosowania indywidualnych pod względem technicznym kanałów produkcyjno-handlowych pomiędzy producentem a odbiorcą. Poza pewnymi specyficznymi przypadkami trudno sobie wyobrazić budowę specjalnych linii energetycznych wykorzystywanych wyłącznie w celu kontaktu z określonym odbiorcą. Energia elektryczna ze źródeł wytwórczych (elektrowni) wprowadzana jest więc do jednego dużego wspólnego systemu elektroenergetycznego, z którego następnie zostaje pobierana przez odbiorców. Tego rodzaju system elektroenergetyczny zintegrowany jest zazwyczaj na poziomie całego kraju.

Krajowy system elektroenergetyczny cechuje duża rozległość terytorialna, ma więc on charakter bardzo rozproszony. Podzielony jest ponadto pod względem organizacyjnym oraz gospodarczym między liczne odrębne podmioty wykorzystujące jego elementy do prowadzenia działalności produkcyjnej oraz handlowej. Należy jednak pamiętać, że pomimo całego rozproszenia organizacyjnego system elektroenergetyczny jest ściśle zintegrowany pod względem technologicznym. Wytwarzanie, przesył oraz dostarczanie energii elektrycznej odbiorcom końcowym stanowią jeden ściśle zintegrowany proces produkcyjno-dystrybucyjny.

Uwarunkowania technologiczne wymagają koordynacji pracy oraz sterowania całością systemu i bilansowania jego działania na różnych szczeblach, począwszy od krajowego, na lokalnym skończywszy. Wyróżniany w nim podsystem dystrybucji (o czym będziemy mówili jeszcze w następnym punkcie), pomimo swej nazwy, nie pełni wyłącznie funkcji związanych z organizacją sprzedaży oraz obsługą klienta. Równie istotną rolę w spółkach obrotu energią odgrywają piony ruchowe (odpowiedzialne za sterowanie pracą systemu elektroenergetycznego w zakresie sieci rozdzielczej zakładu) oraz eksploatacyjne (związane z jego utrzymaniem).

Dane dotyczące potrzeb konsumentów to jedna z podstawowych informacji wejściowych dla organizacji. Ich prognozowanie nie leży być może w głównym nurcie działań związanych ze sterowaniem działalnością podstawową przedsiębiorstwa, stanowi jednak istotną część procesu zaspokojenia potrzeb klienta, który można podzielić na pięć głównych czynności (Muhlemann, Oakland, Lockyer 1995):

– określenie potrzeb prowadzące do ustalenia ich szczegółowych prognoz,
– analiza i integracja prowadzące do ustalenia planów zasobów,
– zaopatrzenie – zapewnienie informacji i środków wejściowych,
– przetworzenie środków w produkt,
– dostarczenie produktu klientowi.

W każdym systemie produkcyjnym istotną rolę odgrywa ponadto sprawna organizacja powiązań umożliwiających efektywny przepływ materiałów i półproduktów. Podstawowe problemy wiążące się z tym zagadnieniem polegają na (Durlik 1995):

– zsynchronizowaniu w czasie wszelkich dostaw, aby skrócić do minimum czas oczekiwania materiału lub półproduktu na wykorzystanie go w procesie wytwórczym, a także na minimalizacji czasu pomiędzy wyprodukowaniem wyrobu a włączeniem do normalnej finalnej eksploatacji,

– zapewnieniu właściwych środków transportu i środków technicznych ułatwiających magazynowanie i wyszukiwanie potrzebnych w danej chwili materiałów, półproduktów lub wyrobów,

– obniżeniu do minimum strat transportowych i magazynowych, zaprojektowaniu takiej struktury przepływu, aby charakteryzowała się ona ograniczeniem dróg transportowych i minimalizacją przeładunków oraz ich pracochłonności.

Na wspomnianą problematykę zaspokajania potrzeb odbiorców oraz organizacji tego procesu nakładają się oczywiście specyficzne cechy systemu elektroenergetycznego, które wpływają na strukturę rynku tego towaru. Z tego punktu widzenia energia elektryczna ma bowiem szereg swoistych właściwości, które muszą zostać wzięte pod uwagę przy konstrukcji odpowiednich rozwiązań rynkowych dla obrotu tym towarem (Zerka 2003; Mielczarski 2000; Lichota 2006). Oto najważniejsze z tych czynników.

1. Udział energii elektrycznej w bilansie energetycznym gospodarki i społeczeństwa, zarówno na skalę światową, jak i lokalną, ma kluczowy charakter. Niezawodność dostaw energii elektrycznej wpływa zatem radykalnie na konkurencyjność gospodarki i poziom życia społeczeństwa. Ewentualne zakłócenia w działaniu rynku tego towaru, a przede wszystkim dostaw elektryczności do końcowych odbiorców, mogą powodować dotkliwe skutki gospodarcze i społeczne o zasięgu nawet międzynarodowym.

2. System elektroenergetyczny zaspokaja potrzeby odbiorców energii dokładnie wtedy, gdy one powstają. Produkcja energii i jej pobór przez odbiorców muszą być bilansowane na bieżąco, w każdej chwili. Niezbędne jest więc uwzględnianie wahań poboru energii w różnych okresach: szczyty dobowe, tygodniowa i roczna zmienność sezonowa. Wymagana jest również natychmiastowa reakcja na zmiany zapotrzebowania spowodowane rozmaitymi czynnikami zewnętrznymi, np. sytuacja pogodowa, jak też wewnętrznymi, np. awarie i zakłócenia w pracy sieci elektroenergetycznej, remonty urządzeń itp. w horyzontach długookresowych.

3. Procesy produkcyjne, z punktu widzenia ich przebiegu w czasie, podzielić można na dyskretne oraz ciągłe. Pierwsze z nich stanowią zestawy operacji umiejscowione logicznie w czasie i przestrzeni, o zmiennej strukturze przystosowanej do charakterystyki ilościowo-jakościowej wytwarzanych produktów. Procesy ciągłe mają z reguły charakter aparaturowy, na trwałe powiązany z urządzeniami produkcyjnymi.

Procesy zachodzące w systemie elektroenergetycznym należą do drugiej grupy. Pobór energii w skali globalnej odbywa się bez przerw, w całodobowym cyklu pracy. Reakcja systemu elektroenergetycznego na wszelkie zmiany zapotrzebowania, nawet o dużej amplitudzie, musi więc odbywać się w czasie rzeczywistym. Przerwy lub brak pełnego pokrycia w dostawach energii dla odbiorców – tzw. energia niedostarczona – to jedna z istotnych zmiennych minimalizowanych w procedurach decyzyjnych dotyczących ruchu i eksploatacji sieci elektroenergetycznej.

4. Istotną cechą systemu elektroenergetycznego jest praktyczna jednoczesność występujących w nim zdarzeń. Szybkość rozchodzenia się zdarzeń w sieci elektroenergetycznej jest rzędu prędkości fali elektromagnetycznej. Każda kilowatogodzina energii, która w danym momencie dociera do odbiorcy, musi zostać wyprodukowana w tej samej chwili, a następnie dostarczona za pośrednictwem sieci przesyłowej i rozdzielczej. Każda zmiana w poborze mocy przez odbiorców musi natychmiast zostać skompensowana przez zmianę mocy wytwarzanej. Nie ma więc żadnych buforów czasowych zarówno na styku z odbiorcami, jak i między operacjami wykonywanymi w obrębie samego systemu elektroenergetycznego.

5. Istotną rolę w synchronizacji procesów produkcyjnych spełniają zapasy zabezpieczające. Zapasy międzyoperacyjne gwarantują zachowanie ciągłości produkcji w warunkach wystąpienia krótkookresowych braków materiałów, narzędzi lub awarii na stanowiskach poprzednich w danym ciągu technologicznym (zapasy gwarancyjne) oraz umożliwiają wyrównywanie okresowych zmian w wydajności pracy (zapasy kompensacyjne) (Durlik 1996). Zapasy gwarancyjne magazynowe umożliwiają synchronizację cyklu zapotrzebowania oraz dostaw.

Niestety brak jest możliwości magazynowania większych ilości energii elektrycznej, która może być przechowywana jedynie na niewielką skalę. Nie da się więc zgromadzić zapasów zabezpieczających dla wyrównania ewentualnych różnic – np. między ilością wytworzonej energii w systemie a zapotrzebowaniem odbiorców. Jednym z poważniejszych wyjątków od tej zasady są elektrownie wodne szczytowo-pompowe. W przypadku nadmiaru energii generowanej w systemie może zostać ona wykorzystana do pompowania wody do zbiorników. Nagromadzona woda wykorzystana zostanie do produkcji energii w czasie niedoborów mocy w systemie. Elektrownie szczytowo-pompowe pełnią bardzo istotną funkcję regulacyjną i interwencyjną w systemie energetycznym. Moc zainstalowana w tego typu źródłach energii jest jednak ograniczona. Potencjal-

nie interesującą dziedziną, jeśli chodzi o magazynowanie energii elektrycznej, mogą być rozwiązania bateryjne, wykorzystywane do wyrównywania cykli pracy odnawialnych źródeł energii (zwłaszcza wiatrowej, słonecznej). Również istotnym elementem mogą stać się tutaj akumulatory samochodów elektrycznych. Tego rodzaju rozwiązania mają jednak charakter przyszłościowy, obecnie nadal jeszcze zakres ich zastosowania należy uznać za niewielki (na lokalną skalę).

6. W krótkim horyzoncie czasowym możliwości zastępowania energii elektrycznej innymi, substytucyjnymi, nośnikami energii są bardzo ograniczone. Problem ten ma szczególnie istotny charakter w przypadku energii wykorzystywanej do zasilania różnego rodzaju urządzeń elektrycznych, używanej do celów oświetleniowych oraz grzewczych. Niewielkie możliwości substytucji stanowią kolejną cechę energii elektrycznej zmuszającą rynek do nieustannego i precyzyjnego równoważenia jej produkcji oraz konsumpcji.

7. Energia elektryczna, w przeciwieństwie do większości innych towarów, wykazuje niemal całkowity brak krótkoterminowej elastyczności cenowej pośród końcowych odbiorców. Fundamentalne znaczenie energii elektrycznej dla społeczeństwa i cywilizacji oraz brak bieżącego przekazu sygnałów ekonomicznych związanych z sytuacją na rynku energii do niemal wszystkich detalicznych użytkowników elektryczności (np. odbiorcy taryfowi w większości rozliczani są z faktycznego zużycia w okresach półrocznych) powodują, że popyt (zwłaszcza krótkookresowy) na nią ma charakter sztywny i w zasadzie nie zmienia się pomimo zmian cen rynkowych. W pewnym ograniczonym zakresie możliwe jest jednak prowadzenie przez przedsiębiorstwa energetyczne oraz odbiorców wspólnych działań mających na celu kształtowanie zapotrzebowania na energię elektryczną przez tychże odbiorców. Rozwiązania tego rodzaju określane są zazwyczaj jako zarządzanie stroną popytową (*Demand Side Management*, DSM), odpowiedź popytowa (*Demand Response*, DR) czy też po prostu aktywny (lub elastyczny) popyt. Jeszcze raz jednak zwróćmy uwagę na ograniczony charakter tego typu działań (przynajmniej w chwili obecnej). Wymagają one nie tylko istnienia specjalnych warunków (np. dany odbiór rzeczywiście musi mieć charakter sterowany), ale także określonych umów między dostawcami a odbiorcami energii, jak również specjalnej infrastruktury telemechaniczno-pomiarowej. W swojej masie popyt na rynku energii elektrycznej traktowany jest w związku z tym jako sztywny zasób, który można (i należy) przewidywać, ale na który (przede wszystkim w krótkim okresie czasu) nie można wpływać.

8. Łatwość monopolizacji rynku energii poprzez wykorzystanie specyficznych cech fizycznych działania systemu elektroenergetycznego.

Wymienione tu cechy energii elektrycznej sprowadzają się w zasadzie do jednej istotnej konkluzji: na rynku energii elektrycznej obowiązuje absolutna i bezwzględna konieczność nieustannego bilansowania jej wytwarzania oraz

wykorzystania. Producenci energii w każdej chwili, jak i w dłuższych odcinkach czasu, nie mogą wytwarzać ani więcej, ani mniej swojego wyrobu niż zużywają odbiorcy. Reguła ta dotyczy rynku globalnego, ale musi, rzecz jasna, w konsekwencji odnosić się również do poszczególnych wytwórców. W przeciwieństwie do większości innych towarów producenci nie mają przecież praktycznych możliwości magazynowania nadmiaru wytworzonej energii. Również w przypadku nadwyżki zapotrzebowania nie ma możliwości jego redukcji przez mechanizmy elastyczności cenowej albo zastąpienia innymi towarami substytucyjnymi. Nie ma także możliwości uzupełnienia niedoborów na rynku zapasami z magazynów. Konstrukcja rynku energii musi więc obejmować mechanizmy równoważenia produkcji i popytu, zarówno w zakresie bieżącym, jak i w dłuższych horyzontach czasowych.

Ponieważ zaspokajanie zapotrzebowania odbiorców odbywa się w sposób ciągły i natychmiastowy, niezbędne jest stałe monitorowanie tego procesu oraz przewidywanie jego możliwych zmian i fluktuacji w różnych zakresach czasowych. W przypadku sekundowych odcinków czasu zmiany procesu obciążenia nie są duże, system może być bilansowany w czasie *quasi*-rzeczywistym poprzez układy samoczynnej regulacji. Dla okresów godzinowych i dobowych możliwe są już jednak znacznie większe wahania poboru energii. Ich bilansowanie wiąże się więc z procedurami uruchamiania i odstawiania jednostek wytwórczych oraz z prowadzeniem wymiany międzysystemowej (Malko 1995). Ponieważ procedury te mają charakter czasochłonny (rozruch bloku elektrowni ze stanu zimnego może trwać nawet kilka dni), informacja o wielkości produkcji danej jednostki wytwórczej musi być określana z pewnym wyprzedzeniem czasowym.

Odbiorcami energii elektrycznej od producentów są w zdecydowanej większości przedsiębiorstwa hurtowego obrotu energią. Zobowiązane są one z góry, z ustalonym wyprzedzeniem czasowym, deklarować wielkość swojego zakupu na pewne ustalone na rynku okresy handlowe. Pamiętać przy tym należy, że nabywca hurtowy w danym okresie handlowym musi zakupić dokładnie tyle energii, ile są w stanie wykorzystać jego odbiorcy, niezależnie od swoich wcześniejszych deklaracji. Tym samym to on przede wszystkim staje przed problemem oszacowania dla celów handlowych wielkości zapotrzebowania na energię swoich odbiorców. Widzimy więc, że prognozowanie popytu na ten specyficzny towar, jakim jest energia elektryczna, nabiera w tych warunkach szczególnie istotnego znaczenia.

Przedsiębiorstwo hurtowego obrotu energią także staje przed problemem nieelastycznego popytu, ponieważ ma ograniczony wpływ na pobór energii przez swoich odbiorców. Ponadto ono również nie może tworzyć zapasów zabezpieczających. Biorąc jednak pod uwagę dużą liczbę odbiorców obsługiwanych przez większość przedsiębiorstw hurtowych, można zazwyczaj przyjąć, że zapotrzebowanie na energię elektryczną to statystyczne zjawisko masowe, które

jesteśmy w stanie prognozować z pewną dokładnością. Z tego samego powodu jednak nie da się określić dokładnych wartości spodziewanego popytu.

W związku z tym rynek energii elektrycznej musi mieć wbudowane pewne mechanizmy z jednej strony przymuszające hurtowych odbiorców do jak najdokładniejszego określania wielkości swojego zapotrzebowania w przyjętych okresach handlowych z ustalonym wyprzedzeniem czasowym, z drugiej strony umożliwiające zakup lub sprzedaż energii w czasie rzeczywistym, by zbilansować faktyczną sprzedaż tych przedsiębiorstw w przypadku nietrafnych przewidywań. Dla przedsiębiorstw działających na rynku energii istotne jest więc nie tylko formułowanie jak najdokładniejszych prognoz zapotrzebowania odbiorców, ale również określanie niepewności tego popytu, co pozwala na szacowanie ryzyka podejmowanych decyzji rynkowych i włączanie ryzyka do procesu decyzyjnego.

Ryzyko wynikające z niepewności popytu stanowi bowiem immanentną część operacji handlowych związanych z energią elektryczną. Niezależnie bowiem od jakichkolwiek czynionych z góry deklaracji wytwórców i nabywców energii, formułowanych w postaci różnorodnych rodzajów umów rynkowych, rzeczywista ilość wyprodukowanej i pobranej energii niemal zawsze będzie od nich w pewnym stopniu odbiegać (Mielczarski 2000). Niemożność określenia z góry dokładnej ilości energii elektrycznej, która zostanie nabyta bądź sprzedana w ramach umów rynkowych w konkretnym obowiązującym na rynku okresie handlowym, jest jedną z najistotniejszych cech odróżniających rynek energii elektrycznej od większości innych rynków towarowych.

Określenie potrzeb klienta w różnych horyzontach czasowych stanowi punkt wyjścia całego procesu produkcyjnego w każdym przedsiębiorstwie wytwórczym. W przypadku systemu elektroenergetycznego szczególnie istotną rolę odgrywa także konfrontacja zapotrzebowania odbiorców energii z możliwościami jego realizacji w danym momencie lub odcinku czasu. Typowe ograniczenia aktualnych zdolności produkcyjnych występują naturalnie również w przypadku przedsiębiorstw elektroenergetyki. Wymienić tutaj możemy (Muhlemann, Oakland, Lockyer 1995):

– parametry instalacji i urządzeń,
– dostępność wyposażenia produkcyjnego,
– dostępność siły roboczej, surowców,
– dostępność gotówki,
– politykę finansowania,
– politykę zaopatrzenia,
– politykę podzlecania prac,
– wymagania techniczne związane z zadaniami,
– podjętą liczbę różnych zadań.

W przypadku specyficznego towaru, jakim jest energia elektryczna, mamy do czynienia jednak z dodatkowymi cechami charakterystycznymi odnoszącymi

się do ścisłego powiązania mechanizmów produkcyjno-handlowych z infrastrukturą techniczną systemu elektroenergetycznego (Mielczarski 2000; Lichota 2006):

– system elektroenergetyczny stanowi w zasadzie jeden wielki obwód elektryczny, przez który energia elektryczna przepływa zgodnie z prawami fizyki, a nie regułami wyznaczanymi przez zawarte wcześniej umowy handlowe między uczestnikami rynku; o rozpływach mocy w systemie elektroenergetycznym decydują przede wszystkim takie elementy, jak struktura sieci energetycznej, jej parametry i zdolności przepustowe oraz chwilowe zapotrzebowanie odbiorców;

– operator zarządzający systemem elektroenergetycznym ma dosyć ograniczone możliwości wymuszenia przepływu mocy od konkretnego wytwórcy do konkretnego odbiorcy; w sterowaniu pracą systemu kieruje się on raczej naczelną zasadą zapewnienia stabilnej transmisji energii elektrycznej zasadniczo od wszystkich producentów do wszystkich odbiorców;

– dodatkowym elementem, który należy uwzględnić, sterując pracą sieci oraz przepływami w niej energii elektrycznej, są tzw. ograniczenia systemowe, takie jak ograniczenia przepustowości linii elektroenergetycznych lub ograniczenia zdolności produkcji w elektrowniach; czynnik ten może mieć znaczący wpływ na ceny energii na rynku.

Jak więc widzimy, w odróżnieniu od innych towarów, energia elektryczna nie może być transportowana od konkretnego wytwórcy do konkretnego odbiorcy. Co więcej, nawet samo określenie konkretnego wytwórcy, którego energię odbiera w danej chwili dany odbiorca, jest po prostu niemożliwe. Wszyscy producenci wprowadzają energię do wspólnej sieci, z której odbierają ją wszyscy odbiorcy. Mechanizmy prowadzenia rozliczeń w ramach konkretnych umów handlowych musi zapewniać rynek energii. Ponadto umowy te, a więc i handel, muszą dotyczyć nie tylko samej energii jako takiej, ale również finansowania elementów dodatkowych związanych z bilansowaniem systemu oraz występującymi ograniczeniami systemowymi.

Specyfika elektryczności jako towaru, silne powiązanie elementów handlowych i technicznych w obrocie energią oraz konieczność stałego równoważenia podaży i popytu sprawiają więc, że rynek energii elektrycznej musi równolegle i w sposób ciągły obsługiwać dwa silnie oddziałujące na siebie procesy – wpływają one w dużym stopniu na konstrukcję mechanizmów handlowych oraz na działania uczestników (Zerka 2003):

1. Proces handlowy – poszczególne podmioty działające na rynku energii konkurują między sobą o sprzedaż lub zakup poszczególnych oferowanych na nim towarów i usług; przede wszystkim dotyczy to, oczywiście, energii elektrycznej, ale także produktów pochodnych, takich jak usługi systemowe związane z przesyłem energii, różnego rodzaju usługi bilansujące zapewniające bezpieczeństwo działania rynku itp. Wynikiem tej konkurencji są różnego

rodzaju kontrakty, które określają i chronią pozycje handlowe uczestników rynku w zawieranych transakcjach.

2. Proces techniczny – zapewnia fizyczną realizację kontraktów zawartych przez podmioty działające na rynku oraz dostaw energii elektrycznej przy uwzględnieniu wszystkich istniejących wymagań technicznych i warunków jakości oraz niezawodności tychże dostaw.

Efektem procesu handlowego na rynku energii są przede wszystkim różnorodne kontrakty określające warunki transakcji gospodarczych między jego uczestnikami. W tej postaci sprzedawana jest i nabywana większość energii, zwłaszcza w obrocie hurtowym. Konieczność określania parametrów transakcji handlowych z pewnym wyprzedzeniem czasowym, która wynika z chwilowego charakteru rynku energii elektrycznej, sprawia, że ta forma ich specyfikacji dominuje. Pozwala to na wcześniejsze ustabilizowanie warunków transakcji, co redukuje ryzyko działających na rynku podmiotów. Kontrakty mogą mieć różnorodny horyzont czasowy, np. wieloletni, roczny, miesięczny, dobowy, a nawet na wybrane godziny doby. Mogą one być zawierane bezpośrednio między producentem a nabywcą, jak również za pośrednictwem różnego rodzaju mechanizmów rynkowych (takich jak giełdy energii).

Zasadniczo kontrakty występujące na rynku energii elektrycznej mają podobny charakter jak te stosowane na innych rynkach towarowych. Pamiętajmy jednak o pewnych, wspominanych już wcześniej, dosyć istotnych różnicach w tej dziedzinie. Przede wszystkim niemożliwe jest dokładne określenie z góry ilości energii, jaka ma zostać nabyta lub sprzedana w okresie, na który zawierany jest kontrakt (Mielczarski 2000). Kontrakty mają więc charakter finansowo-rozliczeniowy, natomiast ich faktyczna realizacja musi zachodzić w procesie technicznym.

To proces techniczny zapewnia bowiem fizyczną dostawę energii w ramach zawartych kontraktów, przy zachowaniu określonych wymagań technicznych działania systemu elektroenergetycznego. Jego koordynacją – planowaniem pracy systemu elektroenergetycznego, rozdzielaniem obciążeń, prowadzeniem bieżącego ruchu w sieci – zajmuje się wyspecjalizowany podmiot. Na bardziej scentralizowanych rynkach energii elementami procesu technicznego zajmuje się zazwyczaj zarządca sieci przesyłowej, tzw. operator systemu przesyłowego. Na rynkach zdecentralizowanych w coraz większym stopniu prowadzone są one bezpośrednio przez uczestników rynku, natomiast operator systemu przesyłowego koordynuje te działania. Usługi w zakresie funkcji bilansujących realizowane są przez mechanizm nazywany rynkiem bilansującym. Prowadzony on jest przez operatora systemu przesyłowego.

Poniżej wymieniamy kilka najważniejszych usługi bilansujących oferowanych w ramach procesu technicznego na rynku energii elektrycznej (Zerka 2003).

1. Bilansowanie energii w systemie elektroenergetycznym, czyli zapewnianie równowagi wytwarzania i zapotrzebowania. Rozumiemy przez to zarówno bilansowanie całego rynku, poprzez odpowiednie planowanie pracy jednostek wytwórczych dla zrównoważenia łącznego popytu, jak i pojedynczych uczestników rynku, dzięki umożliwieniu im zakupu bądź sprzedaży energii elektrycznej w warunkach „ostatniej szansy", na pokrycie indywidualnego niezbilansowania podczas fizycznej realizacji kontraktów. Ponadto w ramach tej usługi wymienić należy prowadzenie rozliczeń finansowych, związanych z niezbilansowaniem. Na nowoczesnych konkurencyjnych rynkach energii tego rodzaju usługi dostarczane są zazwyczaj przez rynek bilansujący.

2. Bilansowanie przepływów energii elektrycznej w sieci przesyłowej, w tym rozwiązywanie problemu aktywnych ograniczeń sieciowych.

3. Bilansowanie usług zapewniających bezpieczeństwo i jakość pracy systemu, a w szczególności bilansowanie rezerw w systemie. Jak już wcześniej powiedzieliśmy, w ustalaniu poziomu produkcji energii w systemie elektroenergetycznym istotną rolę odgrywają prognozy zapotrzebowania. Pojawia się więc problem zapewnienia pokrycia rzeczywistego zapotrzebowania w przypadku zbyt niskich przewidywań popytu lub niespodziewanych wydarzeń powodujących obniżenie zdolności produkcyjnych części wytwórców. Przypomnijmy, że nie ma możliwości stworzenia zapasów gwarancyjnych energii, które mogłyby zostać wykorzystane do pokrycia tego typu niedoborów. Problem ten rozwiązywany jest poprzez utrzymywanie w systemie stałej rezerwy zdolności produkcyjnych (mocy). Wybrane jednostki wytwórcze, niewchodzące do produkcji w danym momencie, utrzymywane są w stanie rezerwy, co umożliwia ich szybki rozruch, a w zamian otrzymują specjalne opłaty za dyspozycyjność. Rezerwa mocy musi więc być na tyle duża, aby zapewnić bezpieczeństwo dostaw energii. Z drugiej jednak strony powinna być ona minimalizowana w celu obniżenia kosztów.

Zadanie rynku energii polega na zapewnieniu swobody obrotu energią elektryczną. Uczestnicy rynku w procesie handlowym dokonują więc transakcji, kierując się własnymi celami ekonomicznymi, takimi jak minimalizacja kosztu zakupu energii w przypadku nabywców czy maksymalizacja zysku ze sprzedaży w przypadku sprzedawców. W tej fazie zazwyczaj nie bierze się pod uwagę kwestii faktycznych dostaw w ramach zawartych kontraktów. Problem jednak polega na tym, że swoboda obrotu energią elektryczną w procesie handlowym niekoniecznie przekłada się na możliwość fizycznej realizacji zawartych transakcji w systemie elektroenergetycznym.

W systemie elektroenergetycznym istnieją bowiem pewne uwarunkowania techniczne związane z konstrukcją jednostek wytwórczych oraz całych elektrowni, a także z procesami fizycznymi zachodzącymi w sieci elektroenergetycznej, które mogą spowodować, że wykonanie określonego zestawu dostaw

energii będzie niewykonalne technicznie. Tego rodzaju uwarunkowania nazywamy ograniczeniami systemowymi.

Ograniczenia systemowe możemy ogólnie określić jako

wymagania techniczne występujące w systemie elektroenergetycznym, zawężające swobodę zmian stanu jednostek wytwórczych oraz wielkości przesyłu energii elektrycznej między obszarami bilansowymi (rozliczeniowymi), których uwzględnienie jest niezbędne dla zapewnienia odpowiednich poziomów jakości i niezawodności pracy systemu (materiały robocze Polskich Sieci Elektroenergetycznych SA, za: Zerka 2003, s. 24).

Ograniczenia systemowe są ważnym czynnikiem, który musi zostać uwzględniony podczas realizacji procesu technicznego na rynku energii. Jest to zjawisko, które w istotny sposób może wpływać na rynkowe ceny energii elektrycznej przy fizycznej realizacji zawartych przez uczestników kontraktów. Z powodu ograniczeń systemowych w ramach mechanizmów rynku bilansującego do pracy niekoniecznie muszą być kierowane jednostki wytwórcze o najniższych kosztach produkcji. Optymalno-kosztowe plany pracy wytwórców wynikające z bilansujących mechanizmów ofertowych rynku sprawdzane są następnie pod kątem uwzględnienia ograniczeń systemowych, co może wymuszać korekty w tym zakresie i sięganie po droższe oferty produkcji, a to wpływa w sposób oczywisty na cenę energii.

Ograniczenia systemowe zazwyczaj dzieli się na trzy podstawowe grupy ze względu na rodzaj wymuszających je uwarunkowań technicznych oraz na miejsce ich występowania w systemie elektroenergetycznym (Zerka 2003):

a) ograniczenia elektrowniane,

b) ograniczenia sieciowe,

c) ograniczenia wynikające z konieczności utrzymania rezerwy mocy w systemie elektroenergetycznym.

Ad a) Ograniczenia elektrowniane związane są w sposób bezpośredni z jednostkami wytwórczymi lub z całymi elektrowniami. Wynikają one z właściwości technicznych poszczególnych jednostek wytwórczych czy też z procedur ruchowych narzucanych przez warunki techniczne pracy jednostek wytwórczych lub całych elektrowni. Ograniczenia te powodują, że nie można w sposób dowolny planować jednostek wytwórczych do pracy, a następnie prowadzić ich ruchu.

W przypadku ograniczeń ze strony jednostek wytwórczych wymienić można ich charakterystyki rozruchowe. Na przykład uruchamianie i odstawianie jednostek wytwórczych w elektrowniach cieplnych związane jest z procesami cieplnymi zachodzącymi w długich okresach. Jednostki wytwórcze mają również określone parametry pracy, np. dotyczące minimalnego czasu pozostawania w danym stanie pracy.

Mówiąc o ograniczeniach ze strony całych elektrowni, możemy wymienić takie parametry, jak minimalna liczba jednostek w danej elektrowni w ruchu,

wymagania ze strony produkcji ciepła, okresy pracy poremontowej, maksymalna liczbę jednostek, jaką można uruchamiać jednocześnie w danej elektrowni itp.

Ad b) Ograniczenia sieciowe wynikają z wymagań w zakresie parametrów lub konfiguracji sieci przesyłowej bądź jej elementów. Wpływają one na możliwości zmiany stanu jednostek wytwórczych, jak również wielkości przesyłu energii elektrycznej między obszarami rozliczeniowymi. Nie każdy bowiem plan dostaw energii wynikający z zawartych umów handlowych między uczestnikami rynku jest następnie technicznie możliwy do realizacji w sieci przesyłowej.

Jako typowe ograniczenia wymienić tu możemy konieczność spełnienia warunków obciążeniowych sieci, warunków napięciowych, warunków zwarciowych, warunków równowagi statycznej i dynamicznej, zakresu regulacji wtórnej mocy, wymagań pewności zasilania obszaru wokół elektrowni.

Ad c) Ograniczenia wynikające z konieczności utrzymania rezerwy mocy w systemie elektroenergetycznym – niezbędne do zachowania rezerwy wirującej (sekundowej, minutowej oraz godzinowej) w jednostkach cieplnych oraz rezerwy mocy w szybko uruchamialnych źródłach szczytowych (elektrownie wodne, gazowe itp.), wykorzystywanej do zapewnienia bezpieczeństwa w zakresie równoważenia zapotrzebowania przez odbiorców.

Problematyka ograniczeń systemowych czy też ogólniej – prowadzenia ruchu w systemie elektroenergetycznym i realizacji usług w procesie technicznym na rynku energii – jest, rzecz jasna, tematem znacznie szerszym i pozostaje poza naszym obszarem zainteresowań w tej pracy. Wspominamy o niej głównie, by zasygnalizować Czytelnikowi pewne specyficzne problemy związane z obrotem energią elektryczną i wpływające na kształt rozwiązań rynkowych w tej dziedzinie. W centrum naszych zainteresowań pozostaje jednak głównie rynek samej energii i to przede wszystkim w zakresie procesu handlowego. Czytelników zainteresowanych szerzej tematyką rynków energii, również w zakresie technicznym, odsyłamy do literatury poświęconej tematyce konstrukcji rynków energii, dla przykładu: Mielczarski 2000; Zerka 2003; Song, Wang 2003; Mielczarski 2005.

1.1.3. Uwarunkowania strukturalne elektroenergetyki

Jak wskazywaliśmy w punkcie 1.1.1, energia elektryczna stała się kluczowym towarem dla rozwoju współczesnej cywilizacji. Powszechność jej wykorzystania zarówno w gospodarce, jak i w życiu codziennym powoduje silne uzależnienie społeczne od tego towaru. Tani i niezawodny dostęp do energii elektrycznej o wysokiej jakości jest warunkiem koniecznym konkurencyjności całej gospodarki oraz zachowania odpowiedniego poziomu życia społeczeństwa. W perspektywie najbliższego ćwierćwiecza, jak wskazują prognozy, znaczenie

energii elektrycznej, a co za tym idzie, uzależnienie cywilizacyjne od tego nośnika energii będzie jeszcze się powiększać.

Nic więc dziwnego, że w ostatnich latach dużo uwagi poświęca się właściwemu ukształtowaniu rynku energii, tak by mógł on coraz sprawniej, taniej i pewniej zaspokajać potrzeby końcowych odbiorców energii elektrycznej. Przy tym wśród podstawowych celów stawianych przed sektorem elektroenergetycznym w tym zakresie można wymienić (Mielczarski 2000):

– jak największą redukcję kosztów dostarczania energii elektrycznej finalnym odbiorcom, zarówno w zakresie przemysłowym, jak i indywidualnym,

– maksymalizację efektywności ekonomicznej sektora elektroenergetycznego,

– ukształtowanie odpowiednich bodźców zachęcających do unowocześniania elektroenergetyki, inwestowania w nowoczesne źródła wytwarzania, infrastrukturę sieciową, ograniczanie emisji zanieczyszczeń i gazów cieplarnianych, poprawę efektywności wykorzystania energii elektrycznej,

– wprowadzenie konkurencji na rynku energii elektrycznej, rozdzielenie świadczenia usług od własności infrastruktury i zapewnienie odbiorcom końcowym możliwości wyboru dostawcy wykorzystywanej przez nich energii,

– wypracowanie rozwiązań prawnych oraz regulacji rynkowych zapewniających konsumentom energii lepszą ochronę przed nadużywaniem siły rynkowej przez przedsiębiorstwa energetyczne,

– poprawę jakości energii dostarczanej końcowym odbiorcom i innych usług oferowanych przez rynek.

Zastanawiając się nad realizacją przedstawionych zadań, musimy wziąć pod uwagę pewne najważniejsze aspekty strukturalne przemysłu elektroenergetycznego oraz ich oddziaływanie na budowę rynku energii. Tradycyjnie, z technicznego punktu widzenia, w systemie elektroenergetycznym wyróżnia się trzy podstawowe podsystemy:

– podsystem wytwórczy – obejmuje elementy systemu elektroenergetycznego związane z wytwarzaniem energii elektrycznej,

– podsystem przesyłowy – obejmuje elementy systemu elektroenergetycznego związane z transportem energii elektrycznej na duże odległości przy wykorzystaniu linii najwyższych napięć (od 220 kV) tworzących sieć przesyłową; energia przekazywana jest od wytwórców do tzw. GPZ (głównych punktów zasilania) poszczególnych sieci dystrybucyjnych,

– podsystem dystrybucyjny – obejmuje część systemu elektroenergetycznego związaną z dostarczaniem energii elektrycznej odbiorcom finalnym przy wykorzystaniu sieci wysokiego napięcia (110 kV), średniego napięcia (głównie 15 i 20 kV) oraz niskiego napięcia, nazywanej siecią dystrybucyjną albo rozdzielczą; energia przekazywana jest z GPZ-ów systemu dystrybucyjnego do poszczególnych odbiorców.

Zaprezentowana struktura systemu elektroenergetycznego uformowała się już niemal od zarania elektroenergetyki w wyniku poszukiwania metody

redukcji kosztów przy przesyle energii na większe odległości i ograniczenia strat sieciowych poprzez zastosowanie wysokiego napięcia w sieciach przesyłowych. Przedstawiony schemat ma naturalnie charakter ogólny i możliwe są pewne drobne odstępstwa od niego. Na przykład podsystem wytwórczy grupuje tylko elektrownie i elektrociepłownie zawodowe (i duże przemysłowe – w Polsce powyżej 0,5 MW). Mniejsze źródła wytwórcze, np. odnawialne źródła energii, podłączone są zazwyczaj bezpośrednio do sieci dystrybucyjnej.

Poszczególne elementy struktury pokrywają się również z najważniejszymi funkcjami pełnionymi przez przedsiębiorstwa elektroenergetyczne, czyli **produkcją** energii elektrycznej, jej **przesyłem** oraz **dystrybucją** do użytkowników. Czwartą podstawową funkcją, która, co prawda, nie znajduje odzwierciedlenia bezpośrednio w strukturze systemu elektroenergetycznego, ale ma chyba dosyć oczywisty charakter, jest **obrót** energią.

Przedstawiona tu struktura systemu elektroenergetycznego w znacznym stopniu również implikuje różne formy uczestnictwa w rynku energii elektrycznej. W tym miejscu zaprezentujemy jedynie wprowadzającą charakterystykę najważniejszych grup podmiotów. Dokładniej do tego zagadnienia wrócimy, analizując struktury współczesnego rynku energii w punkcie 1.2.1. I tak wśród podmiotów biorących udział w rynku energii możemy wyróżnić następujące podstawowe grupy uczestników (UOKiK 2011):

1. Wytwórcy – czyli elektrownie i elektrociepłownie; konwencjonalne, wykorzystujące nieodnawialne źródła energii pierwotnej (oparte na węglu kamiennym, węglu brunatnym, gazie) oraz niekonwencjonalne, wykorzystujące źródła odnawialne (przede wszystkim elektrownie wodne, na biomasę i wiatrowe). Działają jako sprzedawcy energii elektrycznej i regulacyjnych usług systemowych.

2. Spółki obrotu – zajmują się handlem energią kupowaną od wytwórców lub od innych spółek obrotu (zajmujących się obrotem hurtowym) i sprzedają ją odbiorcom końcowym lub innym spółkom. Celem działania przedsiębiorstwa obrotu jest osiąganie zysku poprzez maksymalizację wielkości obrotu oraz różnic pomiędzy cenami sprzedaży i zakupu energii.

3. Operatorzy sieciowi, czyli podmioty będące właścicielami sieci elektroenergetycznej i kierujące jej pracą; wśród nich należy wymienić operatora systemu przesyłowego (OSP) zarządzającego siecią przesyłową oraz operatorów systemów dystrybucyjnych (OSD) zarządzających sieciami dystrybucyjnymi na określonych obszarach geograficznych.

4. Odbiorcy końcowi – odbiorcy przemysłowi kupujący energię na potrzeby prowadzonej przez siebie działalności gospodarczej oraz gospodarstwa domowe (odbiorcy komunalni, rezydencjalni) kupujące energię w celach komunalno-bytowych.

Przedsiębiorstwa elektroenergetyczne mogą pełnić pojedyncze funkcje w jednym z podsystemów systemu elektroenergetycznego. Często jednak

integrują one różne funkcje na różnych jego poziomach. Możemy mówić o **integracji (konsolidacji) pionowej** – czyli łączeniu funkcji pełnionych przez różne poziomy systemu elektroenergetycznego (np. łączenie wytwarzania i dystrybucji energii) albo o **integracji (konsolidacji) poziomej** – czyli łączeniu działań w obrębie określonej funkcji realizowanej na jednym określonym poziomie (np. łączenie kilku elektrowni).

Przez długi czas funkcjonowania elektroenergetyki, do lat dziewięćdziesiątych XX w., panowało przekonanie, że prawidłowe jej działanie może zapewnić tylko silny monopol. Uważano wręcz, że sektor elektroenergetyczny jest naturalnym monopolem, w którym problemy techniczne, technologiczne i inwestycyjne dominują zdecydowanie nad ekonomicznymi. Wśród podstawowych cech przedsiębiorstw elektroenergetycznych z tego okresu należy wymienić (Michalski, Krysta, Lelątko 2004):

– pionowo zintegrowana struktura organizacyjna łącząca różne (zazwyczaj wszystkie) funkcje elektroenergetyki,

– monopol geograficzny na działalność na danym terenie; odbiorcy energii byli „przypisani" do danego dostawcy energii położonego na terenie ich lokalizacji, przy czym monopol ten dotyczył wszystkich funkcji, zarówno dystrybucyjnych, jak i wytwórczych,

– całkowite powiązanie świadczenia usług z własnością infrastruktury; brak rozdzielenia funkcji obrotu energią elektryczną z funkcjami wytwórczymi i transportowymi.

Innymi słowy, przedsiębiorstwa energetyczne były monopolami, które na danym obszarze geograficznym produkowały, dostarczały i sprzedawały energię swoim odbiorcom. Miały one często zasięg krajowych systemów elektroenergetycznych i pozostawały własnością danego państwa. Nawet w Stanach Zjednoczonych, gdzie przemysł elektroenergetyczny należał do prywatnych korporacji, produkcja i przesył energii odbywały się zgodnie ze ścisłymi regułami ustalanymi przez organy administracji państwowej (Mielczarski 2000).

Przyczyny formowania się monopoli w przemyśle elektroenergetycznym związane były z omawianymi w poprzednim punkcie cechami energii elektrycznej jako towaru, jej kluczowym znaczeniem dla odbiorców oraz koniecznością ścisłego powiązania procesów handlowych i technologicznych na rynku energii w zakresie funkcji bilansowania systemu elektroenergetycznego oraz prowadzenia jego ruchu, przy zachowaniu ograniczeń sieciowych. Równie istotną, jeśli nie ważniejszą, rolę odgrywają tutaj pewne właściwości samego przemysłu elektroenergetycznego, o których pamiętać należy nawet, gdy mówimy o bardziej zliberalizowanych i konkurencyjnych rynkach energii. Należą do nich:

– ogromna kosztochłonność budowy nowej infrastruktury wytwórczej i sieciowej w systemie elektroenergetycznym,

– długi okres budowy elektrowni, sieci przesyłowych i rozdzielczych,

– długi okres zwrotu z inwestycji elektroenergetycznych,

– znaczne korzyści wynikające ze skali produkcji energii elektrycznej, zwłaszcza w tradycyjnych źródłach, co sprzyja powstawaniu dużych kompleksów jednostek wytwórczych i utrudnia wchodzenie na rynek małym przedsiębiorstwom elektroenergetycznym,

– znaczna siła rynkowa producentów energii i właścicieli infrastruktury sieciowej/dostawców energii elektrycznej w stosunku do jej finalnych odbiorców i przedsiębiorstw pełniących wyłącznie funkcje handlowe.

Biorąc pod uwagę wspomniane tu powody, należy stwierdzić, że konkurencja w obszarze infrastruktury elektroenergetycznej jest po prostu nieopłacalna. Pomijając już kwestie związane z systemem przesyłowym, nawet w obszarze sieci rozdzielczych po prostu nie opłaca się tworzyć wielu systemów dystrybucyjnych działających w jednej lokalizacji geograficznej. W połączeniu z obowiązującym w poprzednim wieku powiązaniem własności infrastruktury ze świadczeniem usług handlowych w sposób naturalny prowadziło to do tworzenia dużych przedsiębiorstw elektroenergetycznych o znaczącej sile rynkowej, które przekształcały się w struktury monopolistyczne.

Dodatkowym elementem wspomagającym tworzenie monopoli w elektroenergetyce było również przeświadczenie (do dzisiaj często pojawiające się w przypadku rynku energii elektrycznej i pokrewnych rynków innych źródeł energii) o strategicznym znaczeniu przemysłu elektroenergetycznego w sensie politycznym. Pragnienie zapewnienia niezależności i bezpieczeństwa energetycznego gospodarki z jednej strony popychało rządy do tworzenia państwowych monopoli elektroenergetycznych, zaś z drugiej, do bardzo ścisłej regulacji państwowej działania elektroenergetyki, co osłabiało siłę rynkową przedsiębiorstw monopolistycznych w stosunku do innych uczestników rynku, a przede wszystkim końcowych odbiorców energii.

Monopole elektroenergetyczne z ekonomicznego punktu widzenia działały ze wszystkimi wadami nieefektywności, typowymi dla tego typu podmiotów gospodarczych (Dietl, Makowski 2010). Pionowa struktura organizacyjna w powiązaniu z połączeniem funkcji infrastrukturalnych i handlowych powodowały brak konkurencji w obrębie przemysłu. Konsekwencją tej sytuacji był brak przejrzystości kosztów oraz nieefektywność kosztowa i przerosty w zatrudnieniu. Wymienić można ponadto częste praktyki dotowania zarówno wewnątrz samego sektora elektroenergetycznego, jak i między różnymi grupami odbiorców (Paska 2010).

Istotną cechą monopoli elektroenergetycznych była także dominacja zagadnień technicznych i inwestycyjnych nad efektywnością ekonomiczną działania. Negatywnie na efektywność elektroenergetyki wpływała również konieczność pełnienia różnorodnych funkcji społecznych narzucanych przez państwa, które w ten sposób realizowały swoje cele polityczne. Niektóre decyzje modernizacyjne i rozwojowe podejmowane były często z powodów politycznych a nie

ekonomicznych, co skutkowało nieadekwatnymi w stosunku do potrzeb efektami prowadzonych programów inwestycyjnych.

Praktyka monopolu elektroenergetycznego dalece odbiegała więc od sformułowanych na początku bieżącego punktu celów, jakie powinien realizować właściwie ukształtowany rynek energii. Stąd też od początku lat dziewięćdziesiątych ubiegłego wieku wiele państw podjęło usilne działania zmierzające do liberalizacji sektorów elektroenergetycznych oraz do ekonomicznego usprawnienia ich działania poprzez wprowadzenie mechanizmów konkurencyjnych na rynku energii elektrycznej.

Działania te zmierzały w kierunku stworzenia rynku energii, na którym spełniona byłaby jedna fundamentalna zasada: **handel energią i energia elektryczna jako towar są rozdzielone od usług związanych z jej dostawą od sprzedającego do nabywcy**. Warunek ten umożliwia powstanie sytuacji, w której obrót energią i zarządzanie infrastrukturą systemu energetycznego prowadzone mogą być przez różne podmioty i/lub odrębnie wyceniane. Pozwala to również na rozdzielenie funkcji w pionowych strukturach przemysłu, zwłaszcza funkcji wytwarzania i dystrybucji.

Ponadto efektywny rynek energii powinien spełniać pewne dosyć standardowe warunki zapewniające uczciwą konkurencję (Mielczarski 2000), takie jak:

– swobodne ustalanie ceny energii elektrycznej na podstawie mechanizmów rynkowych, działające na zasadzie zrównoważenia popytu i podaży,

– równe prawa podmiotów działających na rynku,

– swobodny dostęp uczestników do rynku oraz jego infrastruktury technicznej, ograniczony wyłącznie warunkami technicznymi lub równymi dla wszystkich warunkami finansowymi.

Konstruując konkurencyjny rynek energii elektrycznej, trzeba naturalnie pamiętać o specyficznych cechach tego towaru omawianych szeroko w poprzednim punkcie tego podrozdziału. Dostawca energii elektrycznej musi wypełnić swoje zobowiązania wobec obsługiwanych końcowych odbiorców. Oznacza to, że jeśli nie zabezpieczył ich zapotrzebowania odpowiednimi własnymi kontraktami zakupu, zmuszony będzie do kupna energii po każdej cenie (Michalski, Krysta, Lelątko 2004). Kolejnym elementem, który musi zostać uwzględniony w budowie mechanizmów konkurencyjnego rynku energii, jest kwestia ścisłego powiązania procesów handlowych i technicznych, którą również omawialiśmy w poprzednim podrozdziale. Rynek energii elektrycznej nie może więc działać wyłącznie w odniesieniu do kryteriów ekonomicznych. Musi uwzględniać określone korekty wynikające z konieczności utrzymania ograniczeń systemowych i równowagi sieci elektroenergetycznej.

Wymienione uwarunkowania powodują, że wprowadzenie pełnej konkurencji w obszarze całego rynku energii należy uznać za niewykonalne. Rozdzielenie energii elektrycznej jako towaru od usług związanych z jej dostawą umożliwia jednak wydzielenie z rynku energii części, w których możliwe jest wprowadze-

nie mechanizmów konkurencyjnych, oraz takich, które muszą pozostać w sferze monopolu.

Nietrudno się, rzecz jasna, domyślić, że w tej drugiej sferze znajdą się obszary rynku silnie związane z infrastrukturą sieci elektroenergetycznej. Jak już wcześniej wskazywaliśmy, kapitałochłonność inwestycji w linie energetyczne powoduje nieopłacalność budowania wielu konkurencyjnych sieci na tym samym obszarze. Ponadto usługi oparte na wykorzystaniu infrastruktury sieciowej są silnie powiązane z procesami technicznymi na rynku energii, ponieważ to ich zadaniem jest realizacja fizycznych dostaw energii zakupionej przez uczestników rynku. Współczesny rynek energii elektrycznej zazwyczaj więc pozostawia sieć przesyłową oraz sieci dystrybucyjne w obrębie monopolu, przy czym sytuacja w obu tych przypadkach jest nieco odmienna.

Zarządzaniem systemem przesyłowym na współczesnym rynku energii zajmuje się zazwyczaj wydzielony podmiot, określany jako operator systemu przesyłowego (OSP). Jego zadanie polega na udostępnianiu zdolności przesyłowych wszystkim uczestnikom rynku na równych prawach. Do obowiązków OSP należy również zapewnienie bezpiecznej i ekonomicznej pracy całego systemu elektroenergetycznego kraju. Dostarcza on opisywanych w poprzednim punkcie usług bilansujących, jak również zapewnia właściwą pracę systemu elektroenergetycznego poprzez wymuszanie spełnienia ograniczeń systemowych i sterowanie pracą jednostek wytwórczych oraz sieci przesyłowej. Zazwyczaj OSP jest jednostką państwową i działa na zasadach silnie regulowanych przez organa administracji państwowej.

Aby ograniczyć siłę rynkową monopolu OSP względem innych uczestników rynku i uniemożliwić nierównoprawne ich traktowanie, udostępnianie zdolności przesyłowych odbywa się na zasadzie ujednoliconych opłat w formie taryf przesyłowych regulowanych przez państwo. Konstrukcja taryf pokrywać musi bieżące koszty działalności, koszty amortyzacji majątku sieciowego oraz określone zyski. Regulacja państwowa służy zapobieżeniu włączaniu do opłat przesyłowych nieuzasadnionych kosztów, związanych np. ze zbędnymi inwestycjami, których celem jest jedynie powiększanie zysków operatora (Mielczarski 2000).

Sieci dystrybucyjne zarządzane są przez operatorów systemów dystrybucyjnych (OSD). Mają one charakter monopolu na danym obszarze geograficznym. Również i w tym przypadku OSD mają obowiązek udostępniać uczestnikom rynku dostęp do sieci rozdzielczej na danym terenie w postaci usług dystrybucyjnych. Wyceniane są one w formie odpowiednich opłat dystrybucyjnych. Podobnie jak OSP, również OSD zajmują się prowadzeniem ruchu i zapewnieniem bezpieczeństwa na obszarze swojej sieci rozdzielczej.

Istnieje pewna poważna różnica między sytuacją operatora systemu przesyłowego i systemów dystrybucyjnych. OSD, w przeciwieństwie do OSP, mogą prowadzić działalność handlową w obrocie energią i zazwyczaj to robią. Funkcję

OSD zazwyczaj pełnią odpowiednie komórki przedsiębiorstw energetycznych, które historycznie dostarczały energię na danym terenie i są na nim właścicielami infrastruktury sieciowej. Kwestie usług dystrybucyjnych muszą być w związku z tym przedmiotem szczególnej uwagi organów regulacyjnych państwa, ponieważ OSD mają tendencję do wykorzystywania swojej siły rynkowej względem innych uczestników rynku poprzez faworyzowanie swoich odbiorców oraz subsydiowanie skrośne, które polega na zaniżaniu opłat za energię przy jednoczesnym zawyżaniu opłat dystrybucyjnych (Mielczarski 2000).

Jak widzimy, zadania nakładane na operatora systemu przesyłowego w zakresie zarządzania pracą całego systemu elektroenergetycznego sprawiają, że na współczesnych rynkach energii organ ten stanowi centralne ogniwo, które spaja je w jedną całość. Bardziej szczegółowo mówiąc, do jego podstawowych obowiązków w tym zakresie należy m.in. (Szczygieł 2001):

– utrzymanie i rozwój systemu przesyłowego w celu zapewnienia bezpieczeństwa dostaw energii elektrycznej,

– prowadzenie ruchu systemu przesyłowego, czyli sterowanie jego pracą w zakresie przepływów energii elektrycznej,

– zagwarantowanie efektywnego, niezawodnego i pewnego funkcjonowania całego systemu elektroenergetycznego,

– zapewnienie równoprawnego dostępu do wszystkich niezbędnych regulacyjnych usług systemowych (usług technicznych),

– uruchamianie, odstawianie i prowadzenie ruchu jednostek wytwórczych,

– równoprawne traktowanie wszystkich użytkowników lub grup użytkowników systemu,

– udostępnianie wszystkim uczestnikom rynku informacji niezbędnych do utrzymania efektywnej i bezpiecznej pracy, koordynacji i rozwoju oraz właściwej współpracy przyłączonego systemu,

– zachowanie poufności informacji handlowych uzyskiwanych w trakcie prowadzenia działalności.

Segmenty związane z samym rynkiem energii elektrycznej jako towaru nie działają bezpośrednio przy wykorzystaniu infrastruktury sieciowej, korzystają z niej tylko poprzez zakup u odpowiednich operatorów usług przesyłowych i dystrybucyjnych. Możliwe jest więc wprowadzenie w tej dziedzinie daleko posuniętych mechanizmów konkurencyjnych. Do obszarów elektroenergetyki, które da się wyodrębnić ze sfery monopolu, zaliczyć możemy przede wszystkim segmenty związane z funkcjami wytwarzania oraz handlu energią elektryczną, przy czym jak zwykle kwestię obrotu energią możemy rozważać na dwóch poziomach: hurtowym i detalicznym.

Konkurencja w wytwarzaniu i obrocie hurtowym związana jest przede wszystkim ze stworzeniem odpowiedniego **hurtowego rynku energii elektrycznej**, na którym konkurują głównie producenci energii i spółki hurtowego obrotu energią, którzy swobodnie kształtują ceny towaru na podstawie kryterium

wzajemnych ofert cenowych. Ograniczenia tej swobody powinny dotyczyć wyłącznie elementów technicznych związanych z zapewnieniem bezpiecznej i stabilnej pracy systemu elektroenergetycznego podczas fizycznej realizacji zawartych umów i transakcji. Uczestnikami hurtowego rynku energii mogą być również bezpośrednio odbiorcy energii elektrycznej, jednakże z reguły dotyczy to tylko największych odbiorców, których zapotrzebowanie jest na tyle duże, że opłaca im się ponieść pewne (dosyć wysokie) koszty udziału w rynku hurtowym.

Rynek hurtowy energii elektrycznej musi przy tym zapewniać wszystkim wytwórcom i nabywcom energii równe prawa. Jego regulacje nie powinny służyć realizacji celów politycznych czy społecznych, związanych np. z zagadnieniami klimatycznymi, emisją zanieczyszczeń czy polityką surowcową w zakresie surowców energetycznych (Mielczarski 2000). Tego rodzaju kwestie muszą być rozwiązywane poza samym rynkiem, np. poprzez określoną politykę podatkową, wykup pozwoleń i certyfikatów emisyjnych.

Konkurencja w obrocie detalicznym, w którym dostawcy energii (spółki obrotu) sprzedają energię elektryczną końcowym odbiorcom. Konkurują oni między sobą, oferując różne warunki dostaw i ceny energii. Pamiętać jednak należy, że obszar sieci dystrybucyjnej na danym terenie stanowi element struktur monopolu naturalnego. Odbiorca, który jest przyłączony do określonej sieci rozdzielczej, nie może zmienić swojej sytuacji w tym zakresie, chyba że zmieni lokalizację geograficzną. Warunkiem niezbędnym do zaistnienia konkurencji w handlu detalicznym energią elektryczną jest więc konieczność udostępnienia sieci dystrybucyjnej przez OSD innym uczestnikom rynku, co określane jest jako **swobodny wybór dostawcy** albo **zasada TPA** (*Third Party Access*), czyli dostępu trzeciej strony (poza operatorem i odbiorcą) do systemu dystrybucyjnego.

Problem stanowi również pojawienie się odpowiedniej liczby alternatywnych dostawców energii elektrycznej dla różnych grup odbiorców. Zasadniczo na znacznie większą skalę swobodny wybór dostawcy dotyczy grupy odbiorców przemysłowych, których jest ograniczona liczba, a przy tym charakteryzują się oni większym poborem energii, są więc znacznie atrakcyjniejszymi partnerami dla niezależnych spółek obrotu energią elektryczną. Sytuacja przedstawia się odmiennie w grupie gospodarstw domowych, czyli odbiorców indywidualnych. Ich cechy charakterystyczne, takie jak znaczna liczba podmiotów oraz niewielki indywidualny odbiór energii, powodują, że sprzedawcy alternatywni często w ogóle nie oferują energii elektrycznej tej grupie odbiorców lub oferują ją po wyższej cenie niż OSD. Z drugiej strony, także odbiorcy w gospodarstwach domowych niechętnie zmieniają dostawcę. Wymaga to podpisania dwóch umów (na usługi dystrybucyjne z OSD oraz na dostawę energii z dostawcą), opłacania dwóch rachunków itp. Odbiorcy indywidualni zazwyczaj uważają za prostsze zawarcie jednej kompleksowej umowy obejmującej jednocześnie obydwa te elementy.

W efekcie bardzo niewielka liczba użytkowników rezydencjalnych korzysta z prawa swobodnego dostawcy energii elektrycznej. Dla przykładu, zgodnie z informacjami Urzędu Regulacji Energetyki (URE), w Polsce, pomimo wprowadzenia w 2007 r. zasady TPA dla wszystkich grup odbiorców, do końca roku 2010 z prawa tego skorzystało zaledwie około 0,01% odbiorców rezydencjalnych (URE 2011). Co prawda, w roku 2011 liczba ta wzrosła dziesięciokrotnie, ale nadal stanowi to tylko około 0,10% uprawnionych z tej grupy (URE 2012).

Z wymienionych tu powodów segment sprzedaży detalicznej energii elektrycznej odbiorcom rezydencjalnym niekorzystającym z zasady TPA jest często wyłączany spod reguł konkurencyjnych. Wprowadza się pojęcie tzw. sprzedawcy z urzędu – „przedsiębiorstwa energetycznego zobowiązanego do świadczenia usług kompleksowych odbiorcom niekorzystającym z prawa wyboru sprzedawcy (zasady TPA)" (UOKiK 2011). Sprzedawcą z urzędu jest z reguły właściciel sieci rozdzielczej, czyli operator systemu dystrybucyjnego. W celu ochrony drobnych odbiorców przed siłą rynkową sprzedawcy z urzędu warunki sprzedaży detalicznej w tym segmencie regulowane są przez państwo. Dla przykładu, w Polsce ceny energii elektrycznej dla odbiorców w gospodarstwach domowych mają charakter urzędowych taryf zatwierdzanych przez Urząd Regulacji Energetyki.

Zwróćmy uwagę na fakt, że rynek energii elektrycznej, nawet o charakterze konkurencyjnym, w wielu aspektach nadal musi być regulowany przez państwo. Znaczenie udziału państwa i skala regulacji w pierwotnych monopolach energetycznych były naturalnie znacznie szersze. Liberalizację energetyki i wprowadzenie zasad konkurencyjnych na rynku energii elektrycznej często określa się wręcz synonimem „proces deregulacji". Ponieważ jednak rynki energii mają pewne elementy pozostające w sferze monopolu bądź narażone są na problemy konkurencji niedoskonałej, rola państwa jako arbitra i regulatora chroniącego słabsze podmioty przed uczestnikami rynku o większej sile lub nawet pozycji dominującej pozostaje nadal bardzo istotna.

W segmencie hurtowym, który na współczesnych rynkach energii ma zdecydowanie bardziej zliberalizowany charakter niż rynek detaliczny, również istnieją poważne zagrożenia związane z wykorzystywaniem siły rynkowej. Wspomniane wcześniej zjawiska, takie jak korzyści skali w produkcji energii elektrycznej oraz wysoka kapitałochłonność inwestycji w infrastrukturę wytwórczą, utrudniają wejście na rynek nowym podmiotom, a sprzyjają dużym producentom. W efekcie rynki energii elektrycznej mają strukturę oligopolu złożonego ze stosunkowo niewielkiej liczby dużych elektrowni.

W idealnej sytuacji konkurencyjnej, jak już wspomnieliśmy, na hurtowym rynku energii elektrycznej producenci i nabywcy energii swobodnie kształtują jej ceny na podstawie ofert cenowych. W praktyce polega to, na ogół, na formowaniu krzywej podażowej w postaci uporządkowanych rosnąco (według proponowanych cen) ofert wytwórców, a następnie skonfrontowaniu jej

z zapotrzebowaniem nabywców. Cena rynkowa energii ustalana jest poprzez przyjęcie ofert produkcyjnych o cenie ofertowej poniżej poziomu akceptacji nabywców. Szczegóły tych operacji mogą się różnić w zależności od mechanizmu rynkowego i konkretnych zastosowanych rozwiązań. Będą one przedmiotem szczegółowej analizy w podrozdziale 1.2.

Tym niemniej bezsporny jest fakt, że uczestnicy o znacznej sile rynkowej łatwo mogą zdestabilizować rynek, wymuszając bardzo poważne skoki cenowe. Brak możliwości magazynowania energii elektrycznej i sztywny popyt zmuszają spółki obrotu do kupowania energii za wszelką cenę w celu pokrycia zapotrzebowania odbiorców. Wycofanie przez dużego producenta części energii z rynku, przy nieelastycznym popycie, może doprowadzić do skokowego wzrostu cen, zwłaszcza w godzinach szczytowych, kiedy inni wytwórcy nie są w stanie pokryć niedoboru. Uzyskane przychody wynikające ze wzrostu cen na rynku mogą z nadwyżką pokryć niższą sprzedaż wytwórcy (Michalski, Krysta, Lelątko 2004). Jest to, rzecz jasna, tylko przykład zagrożeń, jakie mogą powstawać w wyniku zaburzeń konkurencyjności na rynku energii elektrycznej. Analiza i monitorowanie siły rynkowej przedsiębiorstw elektroenergetycznych oraz przeciwdziałanie nadmiernej koncentracji rynku (bądź jego elementów) stanowią więc kolejne ważne zadania urzędów regulacyjnych. Zadania te nabierają szczególnie istotnego znaczenia w świetle obserwowanych w ostatnich latach procesów konsolidacyjnych w sektorach elektroenergetycznych.

Konsolidacja przedsiębiorstw elektroenergetycznych może oczywiście powodować powstanie rozmaitych zagrożeń dla konkurencyjności rynku energii, ale ma również swoje pozytywne strony. Poprzez wykorzystanie efektów skali pozwala na obniżenie kosztów w wytwarzaniu, ale również w obrocie, energią elektryczną. Ponadto większe przedsiębiorstwo łatwiej może udźwignąć wysokie koszty niezbędnych inwestycji i modernizacji w infrastrukturze elektroenergetycznej.

1.2. Mechanizmy ustalania równowagi popytowo-cenowej na chwilowym rynku energii elektrycznej

1.2.1. Struktura konkurencyjnego rynku energii elektrycznej

W prezentowanym podrozdziale zajmiemy się analizą struktury konkurencyjnego rynku energii elektrycznej w segmencie hurtowym. Charakterystyka wszystkich aspektów tego zagadnienia stanowiłaby naturalnie materiał na odrębną książkę, a w większości mają one mniejsze znaczenie dla tematyki naszej pracy. Dlatego jedynie naszkicujemy najistotniejsze zagadnienia z tej dziedziny, tak by dać Czytelnikowi pełniejszy obraz rynku energii elektrycznej

i ułatwić zrozumienie jego działania. Skoncentrujemy się głównie na mechanizmach ustalania równowagi popytowo-cenowej oraz na wpływie niepewności zapotrzebowania na procesy rynkowe. Ponadto przedmiotem naszych zainteresowań, zgodnie z tematem pracy, będzie przede wszystkim rynek krótkoterminowy, transakcji natychmiastowych (*spot*).

Wyróżnia się cztery podstawowe formy rynku energii, które historycznie mogą być traktowane jako kolejne etapy przechodzenia od monopolu elektroenergetycznego do konkurencyjnego rynku energii (Mielczarski 2000; Paska 2010):

1. **Monopol elektroenergetyczny** – pionowo zintegrowane wszystkie funkcje przemysłu elektroenergetycznego, zarówno na poziomie wytwarzania, przesyłu jak i dystrybucji; brak konkurencji na rynku hurtowym jak i detalicznym; odbiorcy zmuszeni są do zakupu energii u dostawcy, do którego sieci są przyłączeni; wysoki poziom regulacji państwowej.

2. **Agencja elektroenergetyczna** – system oparty na jednej niezależnej agencji kupującej energię od wszystkich wytwórców i sprzedającej ją wszystkim dystrybutorom; przykładem może być tu system obowiązujący w Polsce w latach dziewięćdziesiątych XX w.; jeden uprawniony nabywca całej energii (model „*single buyer*"), który ma wyłączność na ustalanie jej ceny i decyzje o rozdziale obciążeń i procesach inwestycyjnych w sieci przesyłowej; zakup energii od wytwórców i sprzedaż dystrybutorom odbywa się z wykorzystaniem kontraktów długoterminowych; tym niemniej w modelu tym następuje podział pionowy struktur elektroenergetyki i wprowadzenie elementów konkurencji wśród wytwórców; brak dostępu stron trzecich do sieci rozdzielczej.

3. **Rynek hurtowy** – na rynku hurtowym wielu niezależnych producentów energii może sprzedawać ją (bezpośrednio lub poprzez mechanizmy pośredniczące) wielu niezależnym spółkom obrotu (lub dużym odbiorcom); swobodna konkurencja na rynku hurtowym ograniczona jest jedynie względami technicznymi i niezbędnym poziomem regulacji; brak konkurencji na rynku detalicznym; odbiorcy detaliczni nie mają prawa swobodnego wyboru dostawcy.

4. **Rynek detaliczny** – pełniejszą konkurencję obserwuje się na rynku hurtowym dzięki szerszemu dostępowi odbiorców, którzy mogą kupować energię bezpośrednio u wytwórców; na rynku detalicznym swobodny wybór dostawcy przez odbiorców, w związku z tym działa na nim wielu hurtowników korzystających z zasady TPA (dostępu stron trzecich do sieci), którzy konkurują między sobą cenami i warunkami dostaw.

Na współczesnych konkurencyjnych rynkach energii elektrycznej hurtowy obrót energią odbywa się zazwyczaj w trzech zasadniczych segmentach: **kontraktowym**, w którym handel realizowany jest w formie kontraktów zawieranych bezpośrednio pomiędzy uczestnikami rynku w obrocie pozagiełdowym (OTC); **giełdowym**, w którym handel prowadzony jest w formie transakcji i kontraktów zawieranych na towarowej giełdzie energii lub za jej

pośrednictwem oraz przy pomocy operatorów handlowo-technicznych; **bilansującym**, gdzie niezbilansowani uczestnicy rynku kupują brakującą lub sprzedają nadmiarową energię, której cena ustalana jest w prowadzonym przez organizatora rynku (operatora systemu przesyłowego) procesie aukcyjnym, z wykorzystaniem specjalnych ofert bilansujących.

Wymienione segmenty rynku hurtowego, zorganizowane być mogą na jeden z trzech następujących sposobów (Szczygieł 2001; Mielczarski 2000):

a) rynek scentralizowany (centralny),

b) rynek giełda – operator systemu przesyłowego,

c) rynek zdecentralizowany (rozproszony).

Ad a) Rynek scentralizowany charakteryzuje się centralnym, najczęściej obligatoryjnym, rynkiem ofertowym (*pool*) prowadzonym przez operatora systemu przesyłowego, który jest jednocześnie operatorem rynku. Oferty sprzedaży energii oraz bilansujące mają charakter zintegrowany. Proces aukcyjny dla każdego ustalonego okresu doby handlowej realizowany jest poprzez ustawienie rosnąco ofert sprzedaży w kolejności cenowej aż do zrównoważenia popytu, w ten sposób wyznacza się cenę rynkową oraz oferty przyjęte do realizacji. Operator rynku odpowiedzialny jest również za proces techniczny, rozdział obciążeń wytwórców i zbilansowanie systemu przy zachowaniu technicznych kryteriów produkcji i przesyłu energii. Innymi słowy, jest to rynek energii o zintegrowanych segmentach giełdowym i bilansującym oraz elementach technicznych, na którym cały handel fizyczną energią prowadzi operator rynku i systemu przesyłowego.

Uczestnicy rynku mogą jednak zawierać między sobą kontrakty finansowe (bezpośrednio, za pośrednictwem instytucji finansowych czy też kontrakty terminowe na działających giełdach). Nie mają przy tym obowiązku informowania operatora rynku o zawartych umowach, kupując fizyczną energię na scentralizowanym rynku zgodnie z chwilową ceną rynkową, a następnie kompensując sobie wzajemnie różnice zgodnie z warunkami zawartych kontraktów. W przypadku zgłoszenia kontraktu dwustronnego operatorowi rynku, jest on uwzględniany przy rozliczeniach. Przepływy finansowe pomiędzy uczestnikami a operatorem rynku dotyczą tylko różnic pomiędzy wolumenem sprzedaży lub zakupu na rynku a wolumenem energii w zgłoszonym kontrakcie.

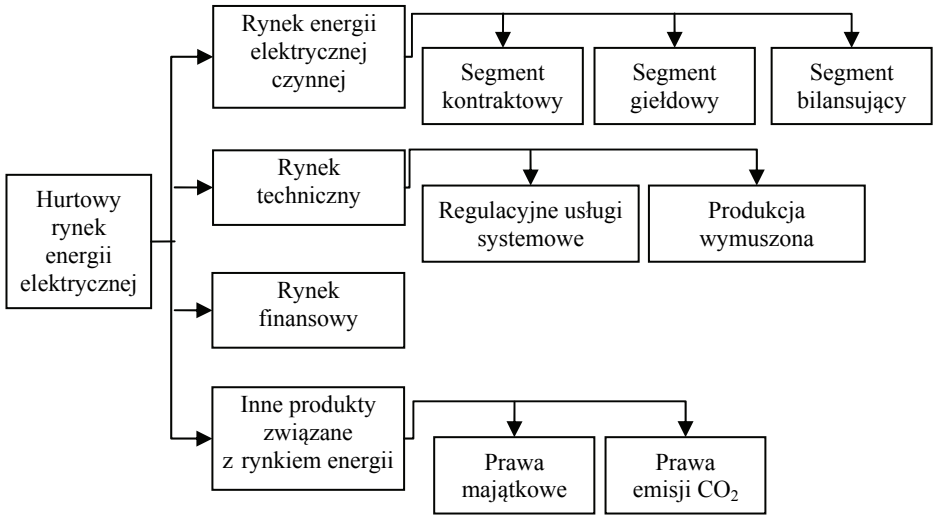
Ad b) Na rynku giełda – operator systemu przesyłowego następuje rozdział segmentu giełdowego i bilansującego, co separuje od siebie (w pewnym stopniu) proces handlowy i techniczny na hurtowym rynku energii. Na rynku funkcjonują więc dwa odrębne krótkoterminowe rynki ofertowe: giełda energii, na której handel realizowany jest na podstawie oferty sprzedaży i zakupu energii, przy założeniu elastycznie cenowego popytu, oraz prowadzony przez OSP, oparty na sztywnym popycie, rynek bilansujący, którego zadanie polega na równoważeniu produkcji energii z zapotrzebowaniem. Operator systemu przesyłowego,

korzystając z wyników działania rynku bilansującego, realizuje również proces techniczny.

Hurtowy rynek energii elektrycznej obejmuje także, jak w poprzednim przypadku, swobodny segment kontraktowy.

Ad c) Funkcje giełdy energii i rynku bilansującego na rynku zdecentralizowanym pozostają bez zmian, jak w poprzednim przypadku. Natomiast struktura rynku uzupełniana jest o operatorów handlowo-technicznych (OHT), którzy zbierają oferty sprzedaży i zakupu energii, scalając je w zbilansowane grafiki obciążeń i zgłaszając operatorowi systemu przesyłowego (lub na rynkach lokalnych operatorowi systemu dystrybucyjnego).

Schemat ogólnej struktury hurtowego rynku energii elektrycznej przedstawiony został na rysunku 1.2.1. Jak widzimy, pod tym pojęciem rozumie się nie tylko rynek samej energii, lecz również kilku dodatkowych towarów i usług wspomagających realizację podstawowych funkcji przedsiębiorstw elektroenergetycznych w obrocie energią. W naszej pracy koncentrujemy się przede wszystkim na sferze obrotu energią, dlatego do pozostałych elementów rynku hurtowego w dalszych rozważaniach odnosić się będziemy w zasadzie sporadycznie.



Rysunek 1.2.1. Ogólna struktura hurtowego rynku energii elektrycznej
Źródło: opracowanie własne na podstawie L. Szczygieł, *Model rynku energii elektrycznej*, [w:] M. Okólski (red.), *Jaki model rynku energii?*, Urząd Regulacji Energetyki, Warszawa 2001

Zwróćmy jeszcze uwagę na określenie „energia czynna" – w przeciwieństwie do „energii biernej" wynikającej z przesuniętej w fazie składowej prądu zmiennego, która generuje się samoczynnie w systemie elektroenergetycznym. Nie analizując szczegółów technicznych, powiedzmy jedynie, że przedmiotem

obrotu na rynku energii jest właśnie energia czynna. Dlatego, jeżeli nie zostanie to wyraźnie inaczej zastrzeżone, używając w naszej pracy określenia „energia", mamy na myśli właśnie energię czynną.

Zgodnie z rysunkiem 1.2.1, w strukturze hurtowego rynku energii elektrycznej wyróżnić możemy cztery rynki składowe: rynek energii elektrycznej czynnej, rynek techniczny, rynek finansowy oraz rynek innych produktów związanych energią. Przyjrzyjmy się krótko każdemu z nich.

Rynek energii elektrycznej czynnej (rzeczywistej). Przedmiotem obrotu na tym rynku jest rzeczywista energia elektryczna o określonych parametrach: ilości, cenie oraz czasie i miejscu dostarczania. Transakcje odbywają się w trzech segmentach: kontraktowym, giełdowym i bilansującym. Pamiętać należy, że chociaż horyzont czasowy umów na dostawę energii może być różny, to zasadniczo rynek ten ma charakter bieżący. Ostateczne fizyczne transakcje związane z całością energii na rynku, prowadzone w segmencie bilansującym, zawierane są w dniu poprzedzającym dobę handlową z określeniem dostaw energii dla unormowanego okresu tej doby (Szczygieł 2001). W Polsce podstawowym okresem handlowym doby jest jedna godzina, stąd też często używa się określenia dobowo-godzinowy rynek energii.

Rynek techniczny. Obsługuje operacje związane z procesem technicznym na rynku energii. Przedmiotem obrotu na tym rynku są rezerwy mocy niezbędnej do równoważenia bieżących zmian zapotrzebowania oraz niektóre usługi systemowe (tzw. regulacyjne usługi systemowe), niezbędne do przesyłu energii zakontraktowanej na rynku. Ponadto obsługuje się na nim obrót energią wytwarzaną przez niektóre źródła w celu spełnienia ograniczeń działania systemu elektroenergetycznego oraz zapewnienia wymogów jakościowych w poszczególnych węzłach sieci (generacja wymuszona).

W początkowych stadiach rozwoju rynku energii operator systemu przesyłowego może kupować od wytwórców niezbędne usługi systemowe na drodze kontraktacji. W bardziej rozwiniętych formach rynku technicznego handel większą częścią usług systemowych odbywa się w formie przetargowej, w postaci bieżącego rynku technicznego, na którym wytwórcy składają oferty na okresy handlowe doby, analogiczne do obowiązujących na rynku energii.

Rynek finansowy. Podobnie jak w przypadku innych rynków towarowych, również na rynku energii elektrycznej mamy do czynienia z obrotem specyficznymi dla niego instrumentami finansowymi. W przeciwieństwie więc do rzeczywistego obrotu energią na rynku finansowym handluje się kontraktami regulującymi należności finansowe związane z partiami energii, dla których określona jest cena i wolumen obrotu, ale które nie są bezpośrednio związane z fizyczną jej dostawą. Rynek ten obejmuje instrumenty pochodne pozwalające uczestnikom rynku na zabezpieczenie się przed ryzykiem związanym z handlem rzeczywistą energią elektryczną jako towarem.

Rynki innych produktów związanych z rynkiem energii. Z działalnością elektroenergetyczną może wiązać się wiele innych produktów, którymi handluje się na rynkach energii. Dla przykładu, przekształcenia sektora elektroenergetycznego związane z redukcją gazów cieplarnianych, zwiększeniem efektywności produkcji energii elektrycznej, przestawienia jej w znacznym stopniu na źródła odnawialne skutkują określonymi formami nacisku na przedsiębiorstwa elektroenergetyczne ze strony organów administracyjnych. Nacisk ten zazwyczaj przyjmuje formę certyfikacji pewnych aspektów działalności. Szereg określonych uprawnień przynoszących wymierne korzyści przedsiębiorstwom elektroenergetycznym ma charakter zbywalny. Wraz z rozwojem rynku energii elektrycznej organizują się więc również rynki, na których obraca się tego typu produktami.

Konkretne produkty pozostające w obrocie zależą już od uregulowań w danym kraju. Przykładem może być tutaj prowadzony przez Towarową Giełdę Energii SA w Warszawie rynek praw majątkowych, na którym handluje się prawami do świadectw pochodzenia dla energii elektrycznej wyprodukowanej w odnawialnych źródłach energii oraz wysokosprawnej kogeneracji (produkcji skojarzonej energii elektrycznej i ciepła). Inny przykład z TGE to rynek uprawnień do emisji CO_2.

Jak już nadmieniliśmy, hurtowy rynek energii elektrycznej rzeczywistej (który dalej będziemy już nazywać po prostu hurtowym rynkiem energii elektrycznej) możemy podzielić na trzy części: kontraktową, giełdową i bilansującą. Poszczególne segmenty będą przedmiotem znacznie bardziej szczegółowych rozważań w kolejnych punktach bieżącego podrozdziału, na razie przyjrzymy się jedynie ich najważniejszym charakterystykom.

Segment kontraktowy przyjmuje na ogół formę kontraktów dwustronnych (bilateralnych) między sprzedawcą (zazwyczaj wytwórcą) a nabywcą energii (zazwyczaj spółką obrotu albo dużym odbiorcą biorącym udział w rynku hurtowym). W związku z tym warunki handlowe tego typu umów, takie jak ceny energii elektrycznej, wolumen, terminy dostaw, ustalane są w wyniku negocjacji między stronami kontraktu. W początkowych fazach wprowadzania rynku energii segment kontraktowy stanowi główny mechanizm hurtowego handlu energią. W miarę rozwoju całości rynku, funkcje kontraktów dwustronnych często przejmowane są jednak w znacznym stopniu przez towarowe kontrakty terminowe w segmencie giełdowym.

Horyzont czasowy kontraktu dwustronnego może zmieniać się od kilku godzin do dni, tygodni, a nawet lat. Warunki kontraktu, takie jak wolumeny dostaw i ceny, mogą być ustalane z góry, w chwili jego zawarcia, na cały okres dostawy, dla poszczególnych okresów handlowych, każdej doby handlowej na rynku energii albo (często przy kontraktach o dłuższym czasie realizacji) określane ogólnie i doprecyzowywane w krótkich terminach przed dostawą. Ceny energii definiowane mogą być również w postaci określonych formuł zależnych od

bieżących cen na rynku. Kontrakty mogą dotyczyć całej doby lub tylko jej wybranego okresu (np. kontrakty typu *Peaks* dla dostaw w okresie szczytowym zapotrzebowania, w ciągu dnia).

Jeżeli kontrakty dwustronne nie są zdefiniowane precyzyjnie z punktu widzenia rynku chwilowego, mają zaś zostać przedstawione operatorowi systemu przesyłowego w celu zaplanowania ich fizycznej realizacji i uwzględnienia w pozycjach stron kontraktu na rynku bilansującym, to musi nastąpić ich grafikowanie, czyli przydzielenie wolumenu energii każdemu okresowi handlowemu dla poszczególnych dób handlowych, zgodnie z zasadami działania rynku bilansującego. Strony kontraktu nie mają oczywiście obowiązku zgłaszać go OSP. W takim jednak przypadku kontrakt przyjmuje charakter wyłącznie finansowy, zabezpieczając (zgodnie z ustalonymi warunkami) pozycje stron na rynkach chwilowych.

Kontrakty służą redukcji ryzyka stron; zapewniają wytwórcom zbyt, a spółkom obrotu zakup energii na pokrycie zapotrzebowania odbiorców, zabezpieczając ich również przed chwilowymi wahaniami cen na rynku. Należy jednak zauważyć, że segment kontraktów dwustronnych generuje również określone poziomy ryzyka, związane przede wszystkim z dwoma źródłami:

1. Ryzyko niewłaściwego określenia warunków kontraktu w chwili jego zawierania; dotyczy zwłaszcza kontraktów o dłuższym okresie trwania. W przypadku spółki obrotu (nabywcy energii) możemy tu wymienić takie elementy, jak ryzyko zmian cen w okresie trwania kontraktu (zarówno na rynku hurtowym, jak i detalicznym) czy też ryzyko fluktuacji wolumenu sprzedaży, czyli znaczących zmian w długoterminowym zapotrzebowaniu odbiorców (Zerka 2001).

2. Ryzyko organiczne, związane z ewentualnym brakiem możliwości realizacji zobowiązań jednej ze stron wynikających z zawartego kontraktu (Zerka 2001). Kontrakty dwustronne stanowią segment obrotu pozagiełdowego (*over-the-counter*), stąd też coraz częściej w stosunku do segmentu kontraktowego hurtowego rynku energii używa się określenia segment OTC. Zazwyczaj problemy z realizacją kontraktów rozstrzygane są na drodze prawa handlowego, nie ma strony trzeciej precyzyjnie nadzorującej ich wykonanie ani systemów zabezpieczeń depozytowych, które występują w kontraktach terminowych na giełdach towarowych energii (lub platformach obrotu energią).

Kolejnym elementem hurtowego rynku energii jest **segment giełdowy**. Spełnia on dwie ważne funkcje, związane z zarządzaniem w obrocie energią.

Giełda energii dostarcza standaryzowanych narzędzi do działań na natychmiastowym rynku energii elektrycznej (*spot*) umożliwiającym handel z bardzo krótkim horyzontem czasowym. Pozwala to na lepsze dopasowanie uczestników rynku energii do zapotrzebowania odbiorców i redukuje ryzyko niezbilansowania. Podstawową formą rynku *spot* na giełdach towarowych energii jest tzw. rynek dnia następnego (RDN), obejmujący transakcje zawierane w dniu poprze-

dzającym dobę fizycznej dostawy (dobę handlową na rynku bilansującym) i dotyczące poszczególnych okresów handlowych tej doby. Giełdy energii mogą również oferować rynki *spot* działające z jeszcze krótszymi wyprzedzeniami czasowymi, w formie rynków dnia bieżącego (RDB).

W zakresie handlu energią elektryczną giełda prowadzi również zazwyczaj rynek towarowych kontraktów terminowych (z fizyczną dostawą towaru), który pozwala na handel z dłuższymi horyzontami czasowymi. Kontrakty giełdowe w odróżnieniu od dwustronnych mają charakter standaryzowany, wykorzystuje się w nich przy tym szereg depozytowych mechanizmów zabezpieczających, które redukują ryzyko potencjalnego niedotrzymania warunków finansowych kontraktu.

Ogólnie mówiąc, handel na giełdowym rynku *spot* ma formę aukcji dwustronnej, tj. z elastycznym popytem. Oferty wytwórców energii ustawiane są w kolejności rosnącej względem cen, odzwierciedlając w ten sposób krzywą podażową. Analogicznie oferty zakupu ustawiane są w kolejności malejącej, zgodnie z krzywą popytową. Parametry oferty występującej na przecięciu obu krzywych wyznaczają cenę równowagi rynku giełdowego, pozwalając na weryfikację strategii ofertowych jego uczestników. Oferty sprzedaży poniżej punktu równowagi oraz oferty zakupu powyżej niego są przyjmowane do realizacji transakcji giełdowych.

W przeciwieństwie do większości innych giełd towarowych, giełda energii elektrycznej musi działać w ścisłym powiązaniu z fizycznymi dostawami energii związanymi z jej przepływami w sieciach elektroenergetycznych, z uwzględnieniem występujących ograniczeń systemowych. Z drugiej jednak strony, zadaniem giełdy jest jedynie obsługa transakcji finansowych. Obsługę fizyczną zobowiązań zawartych na giełdzie musi zapewnić operator systemu przesyłowego. Dlatego po przyjęciu ofert i przeprowadzeniu aukcji giełda przekazuje informacje o przyjętych ofertach zarówno uczestnikom, jak i operatorowi systemu przesyłowego, który realizuje te oferty w miarę możliwości technicznych systemu elektroenergetycznego (Szczygieł 2001). Wystąpienie ewentualnych ograniczeń systemowych nie ma wpływu na rozliczenia finansowe na giełdzie.

Sama giełda jest również uczestnikiem rynku bilansującego. Transakcje rynku *spot*, z samej zasady konstrukcji rynku, tworzą tzw. pozycję zamkniętą, tzn. wolumeny energii zakontraktowanej i dostarczonej są równe.

Giełdy energii prowadzą zazwyczaj działalność nie tylko na samym rynku energii rzeczywistej. Podobnie jak inne giełdy towarowe, obok towarowych kontraktów terminowych rozliczanych przez fizyczną dostawę energii, prowadzić mogą obrót finansowymi instrumentami pochodnymi, które pozwalają na zabezpieczanie transakcji na rynku energii rzeczywistej. Giełdy energii często organizują również transakcje innymi produktami związanymi z energią. Przykładem może być tutaj wspomniany już obrót prawami majątkowymi do świadectw pochodzenia dla energii elektrycznej wyprodukowanej w odnawial-

nych źródłach energii oraz wysokosprawnej kogeneracji albo obrót uprawnieniami do emisji CO_2 prowadzony przez Towarową Giełdę Energii w Warszawie.

Ostatni segment hurtowego rynku energii elektrycznej stanowi **rynek bilansujący**. Ma on nieco odmienny charakter, ponieważ to właśnie na nim wykonywana jest obsługa fizycznej realizacji transakcji finansowych energią zawartych w pozostałych segmentach. Jego zadanie polega na ostatecznym zbilansowaniu wytwarzania i zapotrzebowania całości rynku energii elektrycznej oraz rozliczeniu niezbilansowania poszczególnych jego uczestników. Operatorem rynku bilansującego jest operator systemu przesyłowego, ponieważ to on odpowiada za zrównoważenie systemu elektroenergetycznego kraju oraz na podstawie wyników działania rynku bilansującego prowadzi potem operacje techniczne związane z zaplanowaniem pracy systemu elektroenergetycznego w celu fizycznej realizacji zakontraktowanych dostaw energii.

Technicznie rzecz biorąc, rynek bilansujący ma charakter rynku natychmiastowego *spot*. Procesy handlowe na rynku bilansującym prowadzone są z krótkim wyprzedzeniem czasowym, najdłużej na dzień przed dobą fizycznej realizacji dostaw (dobą handlową). Dotyczą one oddzielnie wszystkich unormowanych okresów handlowych w trakcie tej doby. W Polsce okresami handlowymi są poszczególne godziny doby handlowej, więc rynek bilansujący składa się w zasadzie z 24 oddzielnych rynków dla każdej godziny. W związku z tym dla wygody, zamiast określenia „okres handlowy", używać będziemy pojęcia „godzina".

Dwie funkcje rynku bilansującego implikują dwojaki udział w nim poszczególnych uczestników: aktywny, związany z kształtowaniem ceny na rynku, oraz pasywny, związany z rozliczaniem własnego niezbilansowania. W ramach tej drugiej funkcji uczestnicy rynku zgłaszają operatorowi systemu przesyłowego grafiki swoich portfeli kontraktowych dla każdej godziny doby handlowej. Obejmują one wolumeny energii w kontraktach i transakcjach sprzedaży/zakupu na daną godzinę zawartych przez nich wcześniej w pozostałych segmentach rynku. Po zsumowaniu dają one łączny wolumen transakcji danego uczestnika rynku, dla danej godziny doby handlowej. Jeżeli rzeczywista sprzedaż/zakup o tej godzinie różni się od zgłoszonej (w przypadku producentów pozycja zgłoszona może być korygowana przez operatora systemu przesyłowego z powodów technicznych), różnica ta (określana jako niezbilansowanie) pokrywana jest przez operatora systemu przesyłowego transakcjami na rynku bilansującym, a ich kosztami obciążany zostaje następnie uczestnik rynku.

Na przykład jeżeli na konkretną godzinę odbiorca ma w zgłoszonym grafiku zakontraktowane łącznie za mało energii i jego rzeczywisty pobór jest większy (pozycja krótka), to musi brakującą energię dokupić na rynku bilansującym. Natomiast gdy na tę godzinę zgłoszone kontrakty łącznie przekraczają rzeczywisty pobór odbiorcy (pozycja długa), to posiadana przez niego nadwyżka jest sprzedawana na rynku bilansującym. Proces ten ma charakter obligatoryjny

i niejako automatyczny, tzn. powstanie niezbilansowania automatycznie skutkuje udziałem w transakcjach sprzedaży/zakupu na rynku bilansującym.

Funkcja pierwsza rynku bilansującego związana jest z ustaleniem równowagi popytowo-podażowej całego rynku energii elektrycznej i określeniem ceny, która obowiązuje dla rozliczeń niezbilansowania poszczególnych uczestników. By spełnić tę funkcję, operator systemu przesyłowego wyznacza łączny wolumen energii elektrycznej zakontraktowanej w systemie na daną godzinę doby handlowej, a następnie porównuje ją z zapotrzebowaniem o tej godzinie. Możliwe są trzy sytuacje (Zerka 2003):

– niedokontraktowania w systemie, gdy zakontraktowany wolumen energii jest niższy od wielkości zapotrzebowania; operator w ramach funkcji bilansowania musi zapewnić wzrost produkcji energii elektrycznej ponad zakontraktowany wolumen lub obniżenie zapotrzebowania;

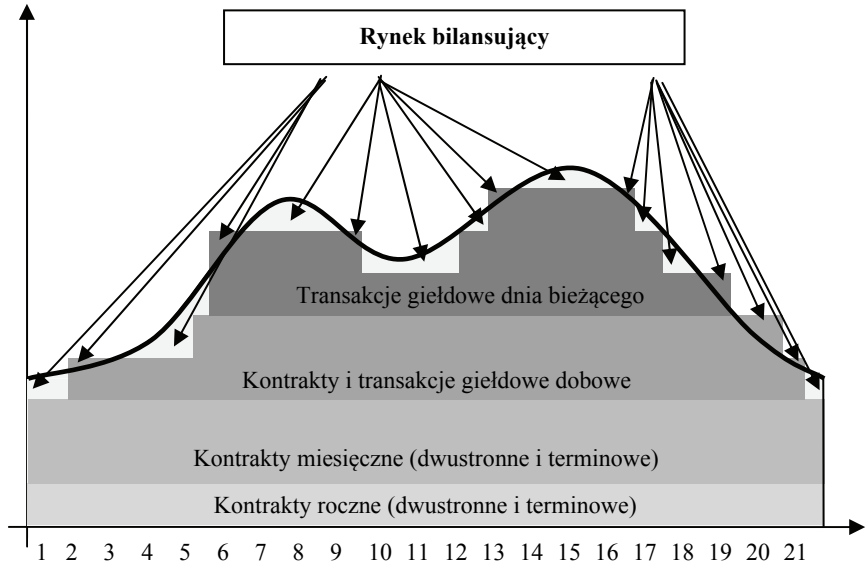
– zbilansowanej kontraktacji, gdy zakontraktowany wolumen jest równy zapotrzebowaniu; nie wymaga to żadnych działań operatora w ramach bilansowania rynku;

– przekontraktowania w systemie, gdy zakontraktowany wolumen jest wyższy od wielkości zapotrzebowania; operator w ramach funkcji bilansowania musi zapewnić obniżenie produkcji energii elektrycznej w stosunku do zakontraktowanego wolumenu lub podwyższenie zapotrzebowania.

Zakup usług bilansujących odbywa się na drodze aukcyjnej. W celu zrównoważenia podaży i popytu wykorzystywane są specjalne oferty bilansujące, głównie ze strony wytwórców, ale również (na rozwiniętych rynkach bilansujących) ze strony nabywców, którzy posiadają sterowalne odbiory końcowe. Aukcje ofert bilansujących mają charakter jednostronny, prowadzone są bowiem dla nieelastycznego zapotrzebowania na energię w krajowym systemie elektroenergetycznym. Oferty porządkowane są niemalejąco względem cen, aż do wyczerpania nadmiaru popytu. Oferta graniczna, na przecięciu krzywej ofertowej i linii popytu, wyznacza równowagę rynku. Na jej podstawie wyznaczana jest cena równowagi (mogą być stosowane tutaj różne podejścia, o czym będziemy mówili w dalszej części tego rozdziału). Oferty usytuowane bliżej na krzywej ofertowej (czyli tańsze od granicznej) są akceptowane i wchodzą do produkcji, znajdujące się dalej (czyli droższe) są odrzucane. Cenę równowagi wykorzystuje się do rozliczeń niezbilansowania poszczególnych uczestników rynku.

Po wstępnym rozdziale produkcji (obciążeń) jednostek wytwórczych, wynikającym z ich portfeli kontraktowych oraz ewentualnych przyjętych ofert bilansujących, operator systemu przesyłowego przeprowadza proces techniczny. Sprawdza i usuwa ewentualne ograniczenia systemowe, co może, jak już wspomnieliśmy, powodować korektę wyników handlowych i cen rozliczeniowych na rynku bilansującym.

Jak więc widzimy, trzy istniejące segmenty hurtowego rynku rzeczywistej energii elektrycznej uzupełniają się nawzajem, oferując różnorodny zestaw instrumentów obrotu energią.



Rysunek 1.2.2. Przykładowy portfel kontraktów i transakcji energią pokrywający zmiany dobowego zapotrzebowania odbiorcy
Źródło: opracowanie własne na podstawie M. Zerka, *Strategie na rynkach energii elektrycznej*, Instytut Doskonalenia Wiedzy o Rynku, Warszawa 2003

Segment kontraktowy, który można rozważać wspólnie (wymiennie) z rynkiem kontraktów terminowych w segmencie giełdowym, obejmuje przede wszystkim narzędzia o dłuższym okresie trwania, giełdowy rynek dnia następnego i dnia bieżącego umożliwia zawieranie bardziej elastycznych transakcji typu *spot*, zaś rynek bilansujący stanowi miejsce obrotu fizyczną energią w czasie rzeczywistym. Dlatego spółki obrotu, by zabezpieczyć zapotrzebowanie swoich odbiorców, powinny być aktywne we wszystkich trzech segmentach rynku energii elektrycznej i utrzymywać portfele różnorodnych instrumentów, które zapewniają zakup odpowiednich ilości energii.

Przykładowy dobowy portfel zakupów zaprezentowany został na rysunku 1.2.2. Narzędzia o dłuższym okresie trwania zabezpieczają podstawową bazę obciążenia spółki i zapewniają długoterminowe bezpieczeństwo handlu i zasilania odbiorców. Bardziej elastyczne narzędzia rynku *spot* pozwalają na lepsze dopasowanie się do krzywej zapotrzebowania odbiorców, których popyt w krótkich horyzontach czasowych można prognozować ze znacznie większą dokładnością. Tym niemniej nigdy nie uda się go przewidzieć z idealną dokładnością, w nieunikniony więc sposób pojawiają się pewne niezbilansowania w postaci krótkich lub długich pozycji kontraktowych spółki. Pokrywane są one odpowiednio zakupem lub sprzedażą „ostatniej szansy" przez spółkę na bilansującym rynku czasu rzeczywistego.

Strategia konstrukcji portfela instrumentów rynkowych, które służą do hurtowego obrotu energią elektryczną w czasie planowanej doby handlowej, zazwyczaj wymaga od uczestnika rynku pewnego zestawu sekwencyjnych decyzji określających jego transakcje (lub oferty transakcyjne) na wielu uporządkowanych czasowo rynkach energii. I tak oto:

1. Uczestnik rynku po analizie portfela posiadanych kontraktów dwustronnych grafikuje je (w porozumieniu z partnerami) na poszczególne okresy handlowe w planowanym dniu, określając wolumeny i ceny kupowanej/sprzedawanej energii, zgodnie z porozumieniami ramowymi poszczególnych umów.

2. Oceniając bardziej precyzyjnie w krótkiej perspektywie czasowej zapotrzebowanie na energię (nabywca) lub zdolności produkcyjne (sprzedawca), mogą dostosować łączny wolumen energii w swoich portfelach kontraktów na każdą godzinę handlową do przewidywanej rzeczywistej wielkości dostaw poprzez transakcje na giełdowym rynku transakcji natychmiastowych (*spot*). Rozwinięty sektor giełdowy oferuje przy tym zazwyczaj kilka rynków *spot*, na których handel odbywa się z różnym (coraz krótszym) wyprzedzeniem czasowym.

3. W kolejnym kroku producenci energii (oraz na bardziej rozwiniętych rynkach niektórzy z odbiorców) mogą wziąć udział w aukcji bilansującej, handlując dodatkowymi kwotami przyrostowymi bądź redukcyjnymi energii elektrycznej. Określają przy tym cenę równowagi na rynku chwilowym. Wszyscy uczestnicy zgłaszają operatorowi systemu przesyłowego zarządzającemu rynkiem bilansującym informacje o swoich umowach handlowych (włączając zakupy i sprzedaż na giełdzie). Ustalają w ten sposób swoje pozycje kontraktowe na rynku bilansującym i określają łączny wolumen energii elektrycznej kupowany/sprzedawany w każdym okresie handlowym. Wystąpienie niezbilansowania danego uczestnika w czasie rzeczywistym skutkuje automatycznie transakcjami na rynku bilansującym.

W kolejnych kilku punktach bieżącego podrozdziału przyjrzymy się bliżej poszczególnym segmentom działania hurtowego rynku energii elektrycznej, zwracając uwagę przede wszystkim na mechanizmy kształtowania się równowagi cenowo-popytowej na rynku oraz kwestie związane z prognozowaniem zapotrzebowania na energię elektryczną i znaczeniem jego niepewności dla procesów rynkowych.

1.2.2. Kontrakty dwustronne

Kontrakty dwustronne (bilateralne) stanowią istotny element hurtowego rynku energii, zwłaszcza w początkowych okresach jego funkcjonowania. Dostarczają one bowiem (między innymi) instrumentów obrotu energią o dłuższych okresach trwania, pozwalając na zmniejszenie lub podział między strony ryzyka

handlowego. Kontrakty dwustronne zabezpieczają więc zazwyczaj podstawową część portfela zakupów energii spółek obrotu. W miarę rozwoju rynku ich funkcje mogą być w znacznym stopniu przejmowane przez towarowe kontrakty terminowe na giełdach energii lub platformach obrotu energią. Giełdy oferują standaryzację i większe bezpieczeństwo obrotu, ich wadą jest jednak to, że korzystanie z usług tego typu pośredników wiąże się ze znacznymi kosztami. Należy również powiedzieć, że na rynku polskim migracja do rynku giełdowego w dużej mierze wymuszona została regulacjami państwowymi dotyczącymi publicznego obrotu energią (tzw. obligo giełdowe).

Oprócz kontraktów dwustronnych w ramach segmentu kontraktowego w Stanach Zjednoczonych i w wielu krajach europejskich (w tym również w Polsce) stosowane mogą być rozwiązania oparte na kontraktach wielostronnych między operatorem systemowym a wytwórcami (lub wytwórcą) z jednej strony a dystrybutorami z drugiej polegające na wieloletnich umowach na dostawę energii elektrycznej. Ich zadaniem jest finansowanie (poprzez zagwarantowanie wytwórcom energii ustalonego poziomu przychodów) projektów inwestycyjnych, w których realizowane są określone specjalne cele, np. służące zapewnieniu bezpieczeństwa energetycznego kraju, rozwijaniu ekologicznych odnawialnych źródeł energii albo zabezpieczeniu zakupów energii dla odbiorców taryfowych.

Skala tego rodzaju porozumień może być różna, od umów z pojedynczym wytwórcą aż do rozwiązań sektorowych obejmujących większość uczestników rynku. Jako przykład podobnych rozwiązań wielostronnych możemy podać kontrakty PPA (*Purchase Power Agreement*) obowiązujące w Stanach Zjednoczonych albo kontrakty długoterminowe (KDT) stosowane do roku 2008 w Polsce.

Mechanizm kontraktów długoterminowych został wprowadzony w Polsce w 1994 r. w celu sfinansowania inwestycji w podsystemie wytwórczym. Zawierane były one przez ówczesnego operatora systemu przesyłowego, Polskie Sieci Elektroenergetyczne SA, z wybranymi w drodze przetargu wytwórcami po cenach wyższych od rynkowych. Wysokość cen ustalana była tak, aby pokryć koszty wynikające z inwestycji. Zakupioną w KDT energię sprzedawano proporcjonalnie pomiędzy operatorów systemów dystrybucyjnych w formie minimum ilości energii (MIE). Kontrakty te miały charakter wieloletnich umów terminowych, a najdłuższe z zawartych miały obowiązywać nawet do 2027 r.

Po wprowadzeniu konkurencyjnego rynku energii KDT utrudniały jego rozwój, ponieważ zamrażały swobodny handel znaczną częścią energii. W początkowej fazie jego funkcjonowania kontrakty długoterminowe obejmowały nawet do 75% rynku. Ponadto miały one wybiórczy charakter, zapewniały korzyści wyłącznie wytwórcom, którzy je zawarli, zakłócając w ten sposób konkurencję na rynku. Po wejściu Polski do Unii Europejskiej zostały uznane przez Komisję Europejską za niedozwoloną pomoc publiczną. W 2008 r.

obowiązujące jeszcze kontrakty długoterminowe zostały rozwiązane przedterminowo na drodze ustawowej, z wykorzystaniem specjalnego systemu odszkodowań.

Obecnie więc na polskim rynku energii elektrycznej segment kontraktowy obrotu energią obejmuje kontrakty dwustronne. Warunki handlowe tych kontraktów, takie jak ceny energii w nich zawarte, wolumeny i terminy dostaw, zastosowane gwarancje ich realizacji, zależą od wyniku negocjacji między stronami i znane są jedynie im.

Okres trwania kontraktu dwustronnego może zmieniać się od kilku godzin do dni, tygodni, a nawet lat. Ogólnie możemy mówić o:

– kontraktach długoterminowych – o kilkuletnich okresach trwania (nie mają nic wspólnego z KDT),

– kontraktach średnioterminowych – o okresie trwania od pół roku do kilku lat, typowe to kontakty roczne,

– kontraktach krótkoterminowych – trwających do pół roku, typowe okresy trwania to kontrakty miesięczne, kwartalne albo tygodniowe,

– kontrakty natychmiastowe (*spot*) – na dzień dostawy, z wyprzedzeniem jedno-, dwudniowym.

Kontrakty mogą dotyczyć całej doby lub tylko jej wybranego pasma godzin (np. kontrakty typu *Peaks* dla dostaw w okresie szczytowym zapotrzebowania w ciągu dnia). Stosuje się również kontrakty na godziny dolin, dolin nocnych, szczytów południowych, popołudniowych. Kontrakty mogą dotyczyć wybranych typów dni: dni roboczych, weekendów.

Aby kontrakt mógł zostać zrealizowany na hurtowym rynku energii elektrycznej, musi mieć określony precyzyjnie wolumen obrotu dla każdej doby dostawy, w poszczególnych jej okresach handlowych (w Polsce w godzinach). Warunki te mogą zostać ustalone z góry, w chwili jego zawarcia, na cały okres realizacji, w postaci sztywnego grafiku dostaw energii. Innym często stosowanym rozwiązaniem jest połączenie określonego harmonogramu z pewnymi możliwościami regulacyjnymi. Wolumeny dostaw w poszczególnych godzinach handlowych każdego dnia realizacji kontraktu specyfikowane są w postaci planowanej podstawy (wolumenu bazowego) oraz zakresu regulacji, czyli przedziału, w którym może zmieniać się faktyczna dostawa od planowanego poziomu. Korekty dostaw dla każdej doby realizacji umowy wykonywane są z odpowiednim wyprzedzeniem, uzgodnionym w warunkach kontraktu. W przypadku niezgłoszenia zmian w harmonogramie wolumenu dostaw obowiązują ustalenia poczynione w grafiku bazowym.

Ceny energii w kontrakcie mogą również mieć charakter cen stałych, ustalanych nominalnie podczas zawarcia kontraktu na każdy okres handlowy wszystkich dób dostawy. Często jednak definiowane są one w postaci określonych formuł dla poszczególnych godzin dostawy, zależnych od ustalonych czynników, takich jak bieżące ceny na rynku, faktyczny wolumen dostawy

(w przypadku zmiennego wolumenu), wartości graniczne ceny. Wynikowa cena w takim przypadku nie jest znana z góry, lecz określa się ją w momencie dostawy poprzez zastosowanie ustalonej formuły.

Pamiętać należy, że jeżeli harmonogramy dostaw w kontraktach dwustronnych nie są zdefiniowane precyzyjnie z punktu widzenia wymagań rynku chwilowego energii elektrycznej, to przed przedstawieniem ich operatorowi systemu przesyłowego do zaplanowania fizycznej realizacji umów i uwzględnienia ich w pozycjach stron kontraktów na rynku bilansującym musi nastąpić ich grafikowanie, czyli przydzielenie wolumenu i cen energii każdemu okresowi handlowemu dla poszczególnych dób handlowych. Oczywiście strony kontraktu nie mają obowiązku zgłaszać go OSP. W takim jednak przypadku kontrakt przyjmuje charakter wyłącznie finansowy, chroniąc (zgodnie z ustalonymi warunkami) pozycje stron na rynkach chwilowych.

Większa część kontraktów to umowy okresowe dotyczące dostaw energii elektrycznej w pewnym przedziale czasu (rok, miesiąc, tydzień). Obok nich na rynku energii elektrycznej mogą funkcjonować, naturalnie, kontrakty jednorazowe dotyczące jednorazowej dostawy energii na określonych w kontrakcie warunkach, np. o określonym wolumenie transakcji oraz jej cenie. Ustalenia formalne związane z zawarciem kontraktu dokonywane są przez jego strony i mogą zależeć od konkretnego przypadku.

Kontrakty okresowe zawierane są zazwyczaj w formie pewnej umowy ramowej, która specyfikuje warunki formalnoprawne związane m.in. z formami i terminami płatności, sposobami rozliczeń, terminami zgłaszania korekt bazy określonego grafiku dostaw, określeniem osób odpowiedzialnych za realizację dostaw energii, zgłoszeń, korekt itd. Towarzyszą jej następnie umowy związane z uszczegółowieniem warunków kontraktu na potrzeby praktycznych dostaw energii. Mogą one przybierać postać kontraktów składowych krótkoterminowych lub porozumień transakcyjnych będących uzupełnieniem umowy ramowej, w których określa się szczegółowe warunki obowiązujące w określonym krótszym okresie, np. dotyczące doprecyzowania ceny i wolumenu dostaw energii elektrycznej. Szczegółowe grafiki dostaw energii elektrycznej w kontrakcie ustalane są często na dzień lub dwa przed dobą dostawy w postaci porozumień natychmiastowych *spot*.

Jak wspomnieliśmy, konkretne warunki kontraktu dwustronnego na dostawę energii elektrycznej zależą od stron zawieranego porozumienia i mogą w poszczególnych przypadkach różnić się między sobą. Istnieje więc wiele możliwych form tego rodzaju umów. Wśród różnorodnych rodzajów kontraktów dwustronnych wymienimy jednak pewną ich formę, tzw. kontrakty różnicowe, którym poświęcimy w bieżącym punkcie nieco więcej uwagi.

Celem kontraktów różnicowych jest zabezpieczenie stron przed zmiennością cen chwilowych na rynku energii poprzez wprowadzenie stałej ceny dostarczanej energii oraz mechanizmu wzajemnych rekompensat za odchylenia

ceny rynkowej od ustalonej. Co istotniejsze dla tematyki naszej pracy, mogą być one często zawierane nie jako umowy fizycznej dostawy energii, ale jako kontrakty finansowe chroniące pozycje stron na chwilowych rynkach energii. Jak pokażemy dalej, pozwalają one bowiem na redukcję ryzyka cenowego, ale jedynie w sytuacji dokładnej znajomości zapotrzebowania na energię. Nawet przy ich zastosowaniu w podejmowanych decyzjach rynkowych nadal musi być uwzględniane ryzyko wynikające z niepewności krótkoterminowego popytu.

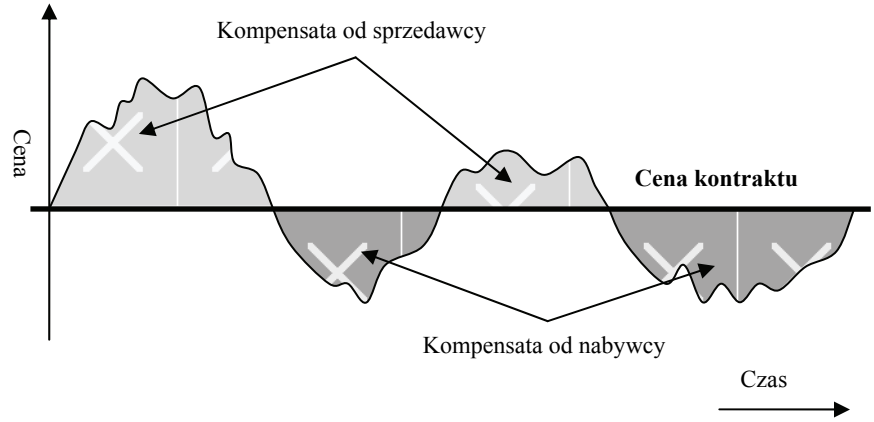
Istnieje oczywiście wiele wariantów konstruowania mechanizmu wyrównawczego cen w kontraktach różnicowych. Wśród najbardziej typowych możemy wymienić:

– kontrakty dwukierunkowe,
– kontrakty jednokierunkowe,
– kontrakty typu maksimum–minimum („kołnierzyk").

Kontrakty dwukierunkowe mają oczywistą i naturalną konstrukcję. Strony kontraktu, ustalając warunki, przyjmują określoną cenę kontraktu i wolumen na każdy okres rozliczeniowy każdego dnia dostawy. Aby zagwarantować wynegocjowaną cenę, rekompensują sobie wzajemnie różnice między nią a chwilową ceną rynkową, tzn.:

– jeżeli cena rynkowa była wyższa od przyjętej w kontrakcie, producent sprzedający energię elektryczną zarobił na zakontraktowanym wolumenie więcej niż wynikałoby to z przyjętej umowy; nadwyżkę tę zwraca nabywcy,

– jeżeli cena rynkowa była niższa od przyjętej w kontrakcie, nabywca zakupił ustalony wolumen, płacąc mniej niż wynikałoby to z umowy; różnicę kwot dopłaca więc sprzedawcy.



Rysunek 1.2.3. Zasada kompensacji ceny chwilowej w kontrakcie dwukierunkowym
Źródło: opracowanie własne na podstawie W. Mielczarski, *Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne*, Agencja Rynku Energii SA i Energoprojekt-Consulting SA, Warszawa 2000, s. 118

Zasadę kompensacji cen dla stron kontraktu zilustrowano również na rysunku 1.2.3. Jak widzimy, rzeczywiście mechanizm ten daje obu stronom gwarancję stałej ceny energii. Oczywiście dokładne rozliczenia w kontrakcie dwustronnym zależą od wielu dodatkowych czynników, m.in. od tego, czy zostanie on zgłoszony operatorowi systemu przesyłowego, czy pozostanie jedynie finansową umową między sprzedawcą a odbiorcą.

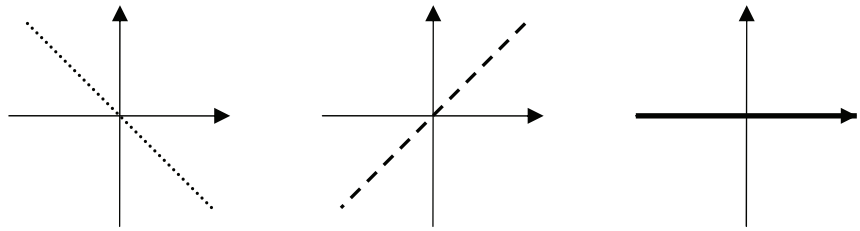
Jak widzimy, kontrakty dwukierunkowe rzeczywiście zapewniają stałą cenę energii elektrycznej, chroniąc sprzedawcę i nabywcę przed zmianami cen chwilowych na rynku. Efekt ten uzyskiwany jest jednak tylko w sytuacji, w której wolumen kontraktu odpowiada dokładnie wielkości dostawy określanej przez zapotrzebowanie odbiorców. Różnice między popytem a kwotą energii w kontrakcie powodują ponowne pojawienie się ryzyka cenowego, spowodowane koniecznością zakupu bądź sprzedaży na rynku chwilowym dodatkowych ilości energii pokrywających niezbilansowanie.

Aby dokładniej przeanalizować przedstawiony problem, przyjrzymy się profilom ryzyka cenowego nabywcy, które wynikają z ewentualnych różnic między ceną rynkową energii elektrycznej a ceną ustaloną w kontrakcie dwukierunkowym. Profile te dla trzech możliwych sytuacji: gdy rzeczywiste zapotrzebowanie energii nabywcy równe jest ustalonemu wolumenowi kontraktu; gdy jego rzeczywiste zapotrzebowanie jest większe od ustalonego wolumenu kontraktu oraz gdy jego rzeczywiste zapotrzebowanie jest mniejsze od ustalonego wolumenu kontraktu, zostały przedstawione, odpowiednio w punktach (a), (b), (c) na rysunku 1.2.4. Oczywiście mają one przykładowy charakter dla wybranych wielkości niezbilansowania.

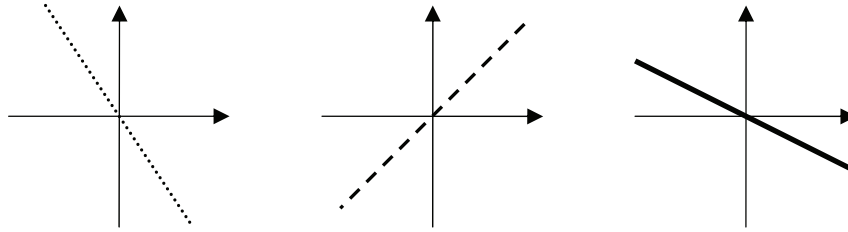
Na wykresach zaprezentowanych na rysunku 1.2.4 zobrazowano zmiany dodatkowego (dodatniego lub ujemnego) przychodu nabywcy energii w kontrakcie dwukierunkowym (oś pionowa) w zależności od (również dodatniej lub ujemnej) różnicy między ceną rynkową a kontraktową (oś pozioma). Obie osie na każdym wykresie przecinają się w punkcie, w którym różnica między ceną rynkową a kontraktową energii wynosi zero (czyli ceny te są sobie równe).

W punkcie (a) na rysunku zaprezentowano sytuację, w której rzeczywiste zapotrzebowanie na energię nabywcy jest równe ustalonemu wolumenowi kontraktu dwukierunkowego. Na pierwszym wykresie przedstawiono zmianę przychodu w przypadku gdyby nabywca całą energię kupował na rynku (czyli nie miał zawartego kontraktu). Jeżeli cena rynkowa w danym okresie jest niska, niższa od ceny w kontrakcie, to różnica między nimi jest ujemna. Korzystając z tej sytuacji, jak widzimy, nabywca osiąga wysokie dodatkowe przychody w stosunku do zakupu z kontraktu.

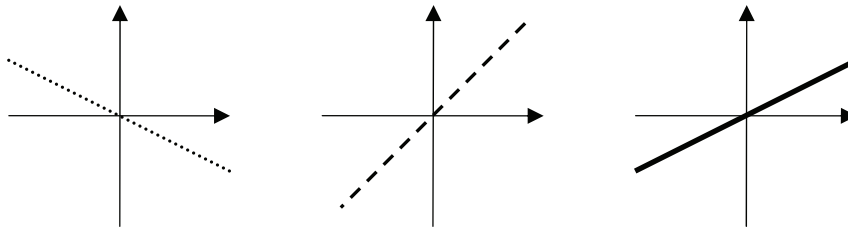
(a) Rzeczywiste zapotrzebowanie równe wolumenowi kontraktu



(b) Rzeczywiste zapotrzebowanie większe od wolumenu kontraktu



(c) Rzeczywiste zapotrzebowanie mniejsze od wolumenu kontraktu



Rysunek 1.2.4. Profile ryzyka cenowego nabywcy w kontrakcie dwukierunkowym przy dokładnym i niedokładnym określeniu wolumenu
Źródło: opracowanie własne na podstawie W. Mielczarski, *Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne*, Agencja Rynku Energii SA i Energoprojekt-Consulting SA, Warszawa 2000, s. 119–120

W miarę wzrostu ceny rynkowej różnica między nią a ceną kontraktową zbliża się do zera, powodując spadek uzyskiwanych przychodów, a następnie staje się dodatnia (gdy cena rynkowa w pewnych okresach handlowych przekroczy cenę kontraktową). Ponieważ cena rynkowa jest wysoka, przy braku kontraktu nabywca zaczyna tracić w stosunku do sytuacji, w której zawarłby go. Prosta, która odzwierciedla dodatkowe przychody z zakupu energii, w tym przypadku jest malejąca. Zauważmy przy tym, że szybkość jej opadania zależy od wolumenu zapotrzebowania nabywcy na energię elektryczną. Oczywiście łączny koszt zakupu równy jest iloczynowi wolumenu zapotrzebowania i różnicy cen, więc zapotrzebowanie na energię odpowiada współczynnikowi kierunkowemu prostej profilu (pomnożonemu przez –1). Im wyższe zapotrzebowanie energii, tym szybciej spada przychód wraz ze wzrostem ceny. Dla niskiego zapotrzebowania nachylenie prostej profilu ryzyka byłoby mniejsze.

Drugi wykres w punkcie (a) na rysunku 1.2.4 odpowiada profilowi ryzyka związanego z rekompensatami nabywcy dla sprzedawcy, które wynikają z zakupu energii w świetle zawartego kontraktu. W tym przypadku naturalnie prosta profilu jest rosnąca, ponieważ przy niskiej cenie rynkowej (niższej od kontraktowej, czyli przy ujemnej różnicy cen) nabywca wskutek działania kontraktu ponosi stratę. Zgodnie z zasadami działania kompensacji w kontrakcie dwustronnym, musi on bowiem zwrócić sprzedawcy kwotę równą wolumenowi kontraktu pomnożonemu przez różnicę cen. Przy wysokich cenach rynkowych skutki kontraktu stałyby się dodatnie, ponieważ w takim przypadku to sprzedawca rekompensuje nadwyżkę cen nabywcy. Tym razem nachylenie prostej dodatkowych przychodów wynikających z kontraktu, dla różnych poziomów cen rynkowych, zdeterminowane jest przez wolumen zakupu z kontraktu.

W sytuacji, w której faktyczne zapotrzebowanie nabywcy na energię równe jest wolumenowi kontraktu, tempo spadku prostej profilu ryzyka przy zakupie energii wyłącznie z rynku, jest zatem dokładnie równe tempu wzrostu prostej profilu ryzyka związanego z kompensacjami wynikającymi z kontraktu. Jeśli weźmiemy pod uwagę obydwa te efekty łącznie, generowane przez nie ryzyko nawzajem się zniesie. Dodatkowe przychody przy każdej różnicy cen równe są dokładnie 0. Jeżeli nabywca energii zarobi dodatkowo, kupując energię na rynku chwilowym przy niskich cenach, to dokładnie taką samą kwotę będzie musiał zrekompensować sprzedawcy i *vice versa*. Profil ryzyka, jak widzimy na trzecim wykresie w punkcie (a) na rysunku 1.2.4, zobrazowany jest za pomocą poziomej prostej pokrywającej się z osią odciętych.

Inaczej przedstawia się sytuacja, gdy rzeczywiste zapotrzebowanie na energię jest większe od wolumenu zawartego kontraktu, co zilustrowano w punkcie (b) na rysunku 1.2.4. Ponieważ wielkości te odpowiadają współczynnikom spadku i wzrostu odpowiednich profili ryzyka, proste na pierwszych dwóch wykresach w tym punkcie nachylone są pod różnymi kątami. Dokładniej mówiąc, prosta na pierwszym wykresie jest mocniej nachylona niż na drugim,

ponieważ współczynnik jej spadku odpowiada faktycznemu popytowi, większemu od wolumenu kontraktu. Przy niskich cenach rynkowych, kupując energię elektryczną na rynku, nabywca zarabia w stosunku do kontraktu kwotę równą swojemu zapotrzebowaniu pomnożonemu przez różnicę cen, natomiast kompensuje sprzedawcy kwotę niższą, równą tylko wolumenowi kontraktu pomnożonemu przez różnicę cen. Przy wysokich cenach rynkowych, dla odmiany, kupując energię elektryczną na rynku, nabywca traci w stosunku do zakupu z kontraktu kwotę równą zapotrzebowaniu pomnożonemu przez różnicę cen, natomiast otrzymuje jako kompensatę od sprzedawcy kwotę równą wolumenowi kontraktu pomnożonemu przez różnicę cen. Profile ryzyka nie znoszą się więc w pełni.

Łączny profil dla tego przypadku przedstawiony jest za pomocą opadającej prostej pokazanej na trzecim wykresie w punkcie (b). Współczynnik szybkości jej spadku równy jest różnicy między współczynnikami kierunkowymi prostych na pierwszych dwóch wykresach. Innymi słowy, jest on określany przez nadwyżkę rzeczywistego popytu nad wielkością zamówioną w kontrakcie. Jeżeli ceny rynkowe w danym okresie handlowym są niższe od cen kontraktowych, nabywca kupuje dodatkową energię na rynku bilansującym, uzyskując dzięki niskim cenom dodatkowy przychód w stosunku do ewentualnego jej zakupu w kontrakcie. Jeżeli w pewnych okresach ceny rynkowe są wyższe od gwarantowanych przez kontrakt, zakup na rynku nadwyżki niezbędnej do pokrycia zapotrzebowania przynosi, w porównaniu z zakupem z kontraktu, pewną stratę.

W punkcie (c) na rysunku 1.2.4 zobrazowano z kolei sytuację odwrotną, w której rzeczywiste zapotrzebowanie na energię elektryczną jest mniejsze od wolumenu kontraktu zawartego przez nabywcę. Oznacza to, że tym razem prosta na pierwszym wykresie opada wolniej niż rośnie prosta na drugim wykresie. Przy niskich cenach rynkowych nabywca kupujący energię elektryczną na rynku zarabia w stosunku do kontraktu kwotę równą swojemu zapotrzebowaniu pomnożonemu przez różnicę cen, natomiast kompensuje sprzedawcy kwotę większą, równą wolumenowi kontraktu pomnożonemu przez różnicę cen. Na tej samej zasadzie przy wysokich cenach, kiedy kompensatę wypłaca sprzedawca, jest ona większa niż koszty zakupu energii na rynku. Ponownie więc mechanizm kompensacji w kontrakcie dwustronnym nie daje nabywcy pełnej gwarancji stałej ceny.

Podsumowując, należy stwierdzić, że kontrakty dwustronne stanowią narzędzie skutecznie likwidujące ryzyko zmian krótkoterminowych cen energii elektrycznej na rynku chwilowym w sytuacji, w której nabywca jest w stanie precyzyjnie określić swoje zapotrzebowanie na energię. W przypadku błędów w ocenie zapotrzebowania przychody nabywcy nadal pozostają w pewnym stopniu eksponowane na zmiany ceny rynkowej. Należy jednak zwrócić uwagę, że poziom tego ryzyka jest znacznie niższy niż w przypadku rezygnacji z zawarcia kontraktu, dotyczy bowiem konieczności zakupu lub sprzedaży różnicy niezbędnej do zbilansowania zapotrzebowania nabywcy.

Drugim powszechnie spotykanym typem kontraktów różnicowych są **kontrakty jednokierunkowe**. Podobnie jak w przypadku kontraktów dwukierun-

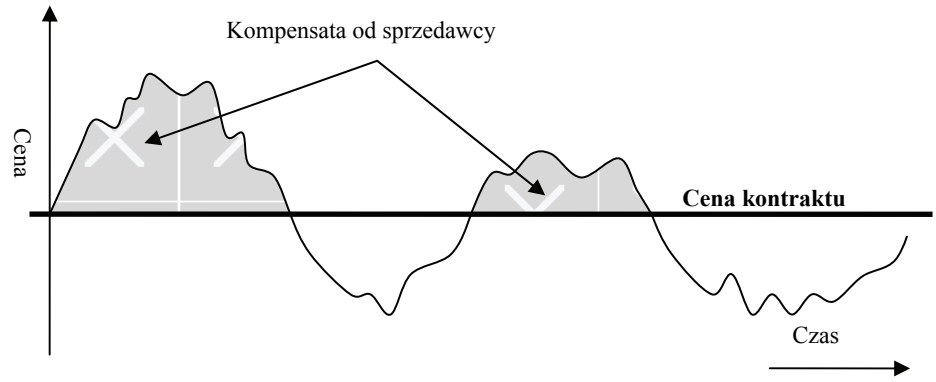
kowych, obie strony przyjmują określoną cenę kontraktu i wolumen na każdy okres rozliczeniowy każdego dnia dostawy. Zasady wzajemnej kompensacji w przypadku odchyleń ceny rynkowej od ustalonej w kontrakcie są jednak w tym przypadku nieco odmienne:

– nabywca płaci sprzedającemu opłatę stałą; zazwyczaj jest ona zależna od wolumenu zakontraktowanej energii, ceny kontraktu i okresu trwania kontraktu,

– jeżeli cena rynkowa była wyższa od przyjętej w kontrakcie, sprzedawca kompensuje nabywcy uzyskaną nadwyżkę,

– jeżeli cena rynkowa była niższa od przyjętej w kontrakcie, nie ma kompensaty dla sprzedawcy.

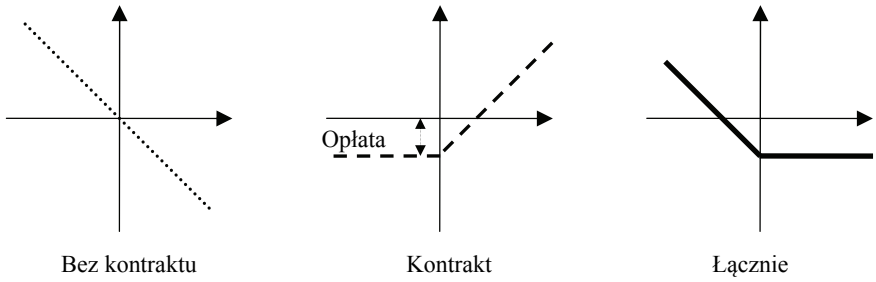
Zasadę kompensacji cenowych dla obu stron zawartego kontraktu jednostronnego zilustrowano na rysunku 1.2.5. Podobnie jak w przypadku kompensacji dwukierunkowych, również dla kontraktów jednokierunkowych musimy rozważyć wpływ niepewności zapotrzebowania na energię. Odpowiednie profile ryzyka dla dodatkowego przychodu związanego z odchyleniem ceny rynkowej względem przyjętej w kontrakcie zostały zaprezentowane na rysunku 1.2.6.



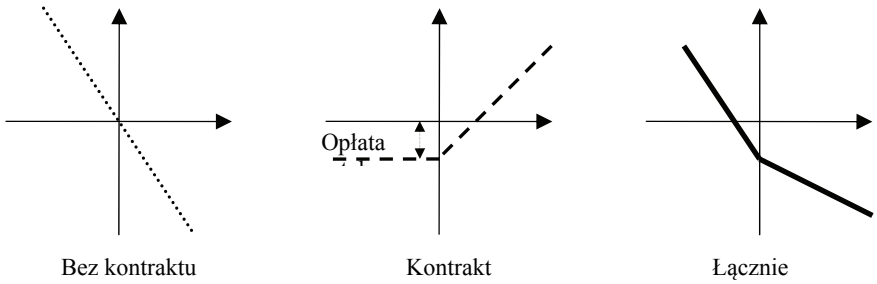
Rysunek 1.2.5. Zasada kompensacji ceny chwilowej w kontrakcie jednokierunkowym
Źródło: opracowanie własne na podstawie W. Mielczarski, *Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne*, Agencja Rynku Energii SA i Energoprojekt-Consulting SA, Warszawa 2000, s. 124

Sposób rozumowania podczas konstrukcji profili ryzyka jest analogiczny jak przy kontraktach dwustronnych, dlatego nie będziemy powtarzali ich szczegółowej analizy, skupiając się tylko pokrótce na wynikających z niej wnioskach. Zauważmy, że nawet w sytuacji, w której zapotrzebowanie na energię nabywcy równe jest wolumenowi kontraktu (punkt (a) na rysunku 1.2.6), zawarty kontrakt jednostronny nie stabilizuje w pełni poziomu cenowego. Przyjrzyjmy się finalnemu profilowi ryzyka ponoszonego przez nabywcę na trzecim wykresie w tym punkcie.

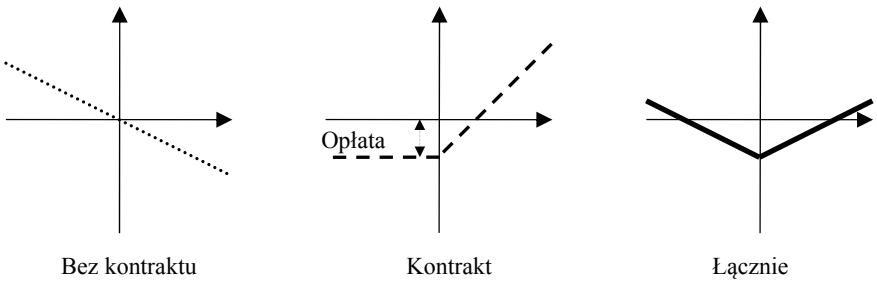
(a) Rzeczywiste zapotrzebowanie równe wolumenowi kontraktu



(b) Rzeczywiste zapotrzebowanie większe od wolumenu kontraktu



(c) Rzeczywiste zapotrzebowanie mniejsze od wolumenu kontraktu



Rysunek 1.2.6. Profile ryzyka cenowego nabywcy w kontrakcie jednokierunkowym przy dokładnym i niedokładnym określeniu wolumenu
Źródło: opracowanie własne na podstawie W. Mielczarski, *Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne*, Agencja Rynku Energii SA i Energoprojekt-Consulting SA, Warszawa 2000, s. 124

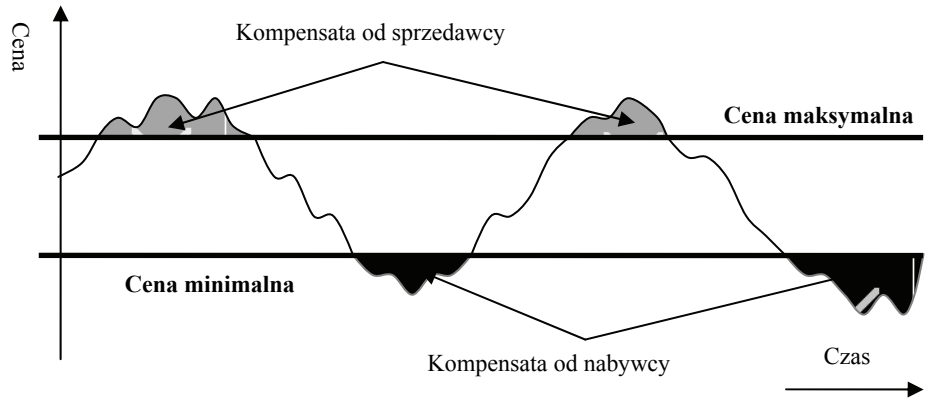
Przy niskich cenach rynkowych, znacznie niższych od ceny ustalonej w kontrakcie, nabywca, kupując energię przez rynek, zyskuje w stosunku do swoich zobowiązań kontraktowych, ponieważ nie musi kompensować sprzedawcy dodatkowych przychodów uzyskanych z tego powodu. Jeżeli jednak w pewnych okresach rozliczeniowych na rynku cena rynkowa energii elektrycznej będzie tylko nieznacznie niższa od kontraktowej, nabywca zacznie tracić, ponieważ przychody uzyskane z zakupu po niższych cenach nie będą mu gwarantować pokrycia stałej opłaty określonej w kontrakcie. Wreszcie, w godzinach, w których cena rynkowa będzie wyższa od kontraktowej, czyli ich różnica stanie się dodatnia, straty nabywcy przestaną rosnąć, stabilizując się na określonym poziomie odpowiadającym opłacie stałej. Widzimy więc, że kontrakt jednostronny nie jest instrumentem, którego główny cel polega na zagwarantowaniu ceny energii na ustalonym poziomie, lecz formą zabezpieczenia, która pozwala nabywcy spekulować w nadziei na niskie ceny rynkowe, natomiast przy wysokim ich poziomie, ograniczyć ponoszone straty.

Oczywiście w przypadku rozbieżności między faktycznym zapotrzebowaniem nabywcy a wolumenem zawartego kontraktu jednostronnego pojawia się dodatkowe ryzyko związane z tym niezbilansowaniem. Kompensacje zawarte w kontrakcie nie w pełni gwarantują ograniczenie strat odbiorcy energii w przypadku wysokiego poziomu cen rynkowych. W sytuacji, w której rzeczywiste zapotrzebowanie jest większe od wolumenu kontraktu, straty nabywcy będą rosły proporcjonalnie do wielkości niedoboru, wraz ze wzrostem ceny rynkowej, co widzimy na łącznym profilu ryzyka przedstawionym w punkcie (b). Jeżeli natomiast rzeczywiste zapotrzebowanie okazałoby się mniejsze od wolumenu kontraktu, wysokie ceny rynkowe mogłyby oznaczać nawet dodatkowy zysk nabywcy, związany ze sprzedażą nadwyżek energii na rynku chwilowym.

Trzecim z powszechnie występujących na rynku energii elektrycznej rodzajów kontraktów różnicowych, są **kontrakty typu maksimum–minimum** (określane również jako **kontrakty typu „kołnierzyk”**). Zadaniem tego rodzaju umów jest utrzymanie ceny energii w ustalonym z góry korytarzu cenowym. Obie strony kontraktu na każdy okres rozliczeniowy, każdego dnia dostawy, obok wolumenu energii, określają także cenę minimalną i maksymalną. Zasady wzajemnej kompensacji są w tym przypadku dosyć oczywiste:

– jeżeli cena rynkowa jest wyższa od ceny maksymalnej przyjętej w kontrakcie, sprzedawca kompensuje nabywcy uzyskaną nadwyżkę,

– jeżeli cena rynkowa jest niższa od ustalonej w kontrakcie ceny minimalnej, nabywca kompensuje uzyskaną nadwyżkę sprzedawcy.



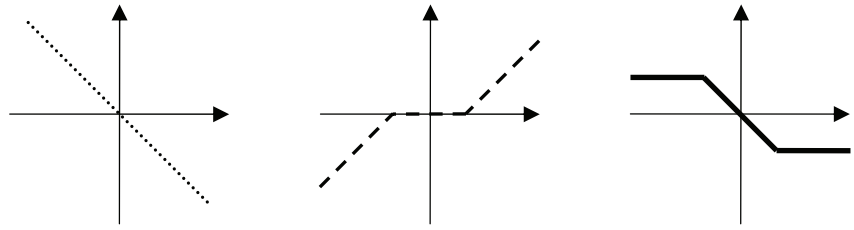
Rysunek 1.2.7. Zasada kompensacji ceny chwilowej w kontrakcie typu minimum–maksimum
Źródło: opracowanie własne na podstawie W. Mielczarski, *Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne*, Agencja Rynku Energii SA i Energoprojekt-Consulting SA, Warszawa 2000, s. 125

Ideę kompensacji cenowych dla obu stron zawartego kontraktu typu maksimum–minimum zilustrowano na rysunku 1.2.7.

Podobnie jak w pozostałych przypadkach kontrakty typu maksimum–minimum zapewniają zawierającym je stronom utrzymanie ceny energii elektrycznej w przyjętym przez nie korytarzu cenowym, pod warunkiem ustalenia wolumenu kontraktu dokładnie na poziomie rzeczywistej dostawy energii do nabywcy. Przyjrzyjmy się więc pokrótce profilom ryzyka nabywcy dla dodatkowego przychodu związanego z odchyleniem ceny rynkowej energii od cen określonych w kontrakcie, w kontekście niepewności jego zapotrzebowania. Profile te dla kontraktów typu maksimum–minimum zostały przedstawione na rysunku 1.2.8.

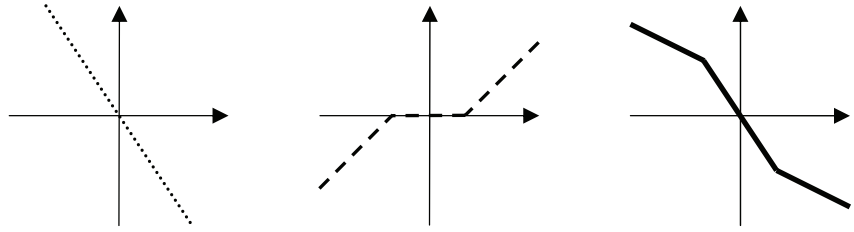
Sposób rozumowania podczas konstrukcji profili cenowych ryzyka, również i w przypadku kontraktów typu minimum–maksimum, jest bardzo zbliżony do postępowania przy kontraktach dwustronnych, dlatego nie będziemy przedstawiali jego szczegółowej analizy, ponownie skupiając się na wnioskach wynikających z finalnego (łącznego) profilu. Jedyna poważniejsza różnica wynika z faktu, że w kontraktach typu minimum–maksimum nie mamy do czynienia z jedną ceną kontraktową energii, ale z ceną maksymalną i minimalną. W związku z tym w profilach ryzyka będziemy badać zmiany przychodów nabywcy dla różnicy między chwilową ceną rynkową a średnią obu cen kontraktowych. Innymi słowy, punkt odniesienia, położony na przecięciu osi na wykresach znajdujących się na rysunku 1.2.8, wyznaczany jest więc tym razem przez środek przyjętego przez strony kontraktu korytarza cenowego.

(a) Rzeczywiste zapotrzebowanie równe wolumenowi kontraktu



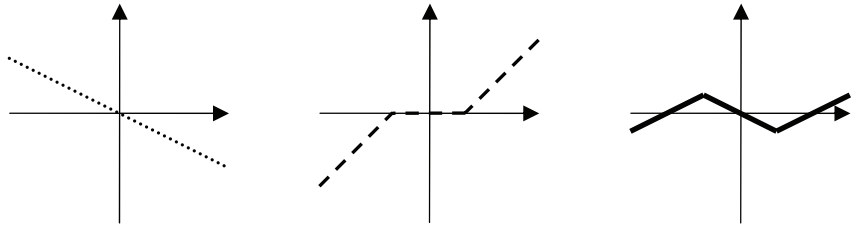
Bez kontraktu Kontrakt Łącznie

(b) Rzeczywiste zapotrzebowanie większe od wolumenu kontraktu



Bez kontraktu Kontrakt Łącznie

(c) Rzeczywiste zapotrzebowanie mniejsze od wolumenu kontraktu



Bez kontraktu Kontrakt Łącznie

Rysunek 1.2.8. Profile ryzyka cenowego nabywcy w kontrakcie maksimum–minimum przy dokładnym i niedokładnym określeniu wolumenu
Źródło: opracowanie własne na podstawie W. Mielczarski, *Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne*, Agencja Rynku Energii SA i Energoprojekt-Consulting SA, Warszawa 2000, s. 125

Jeżeli spojrzymy na finalny profil ryzyka cenowego nabywcy energii (trzeci wykres) na punkt (a), czyli na sytuację, w której odbierana dostawa jest równa wolumenowi zawartego kontraktu, to widzimy, że kontrakt typu maksimum–minimum w tych warunkach stabilizuje ryzyko cenowe kupującego w granicach ustalonego wcześniej przez strony korytarza cenowego. W jego obrębie przychody nabywcy spadają wraz ze wzrostem cen rynkowych, w tempie wyznaczanym przez zapotrzebowanie/wolumen kontraktu. Jednocześnie dla bardzo niskich albo bardzo wysokich cen rynkowych dodatkowy wynikający z nich przychód nie zmienia się, ponieważ wzajemne kompensaty powodują jego stabilizację na wyznaczonej cenie minimalnej i maksymalnej kontraktu.

Jeżeli rzeczywiste zapotrzebowanie na energię jest wyższe albo niższe od wolumenu zawartego kontraktu, dodatkowe ryzyko związane z niezbilansowaniem powoduje, że ustalone w kontrakcie kompensacje nie w pełni gwarantują ograniczenie strat nabywcy w przypadku wysokiego poziomu cen rynkowych czy też strat sprzedawcy przy niskich cenach. Odpowiednie profile ryzyka dla przychodów nabywcy zostały przedstawione w punkcie (b) i (c) na rysunku 1.2.8.

1.2.3. Giełda energii

1.2.3.1. Ogólna charakterystyka

Segment giełdowy, jak wspomnieliśmy w punkcie 1.1.1, pełni dwie istotne funkcje związane z konstruowaniem przez uczestników rynku portfela kontraktów na zakup/sprzedaż energii elektrycznej.

Przede wszystkim giełda dostarcza zestawu instrumentów umożliwiających handel energią z bardzo krótkim horyzontem czasowym (jedno- lub dwudniowym) – w ten sposób tworzy się standaryzowany pod względem formy obrotu rynek transakcji natychmiastowych (*spot*) stanowiących element pośredni pomiędzy segmentem kontraktów dwustronnych a działającym w czasie rzeczywistym rynkiem bilansującym. W krótszej perspektywie czasowej nabywcy na hurtowym rynku energii mogą lepiej dostosować swoje propozycje ofertowe do rzeczywistego zapotrzebowania końcowych odbiorców, zaś sprzedawcy – do swoich aktualnych możliwości produkcyjnych, redukując dzięki temu ryzyko poważnych wielkości niezbilansowania i skalę udziału w rynku bieżącym (bilansującym). We wczesnych stadiach rozwoju, rynek *spot* przyjmuje zazwyczaj formę aukcji z dostawą w dniu następnym, w późniejszym okresie wprowadzane mogą być jeszcze krótsze wyprzedzenia czasowe.

Giełdy energii elektrycznej prowadzą również zazwyczaj rynki towarowych kontraktów terminowych, które umożliwiają handel z dłuższymi horyzontami czasowymi. Kontrakty giełdowe w odróżnieniu od dwustronnych mają charakter

standaryzowany i wykorzystują różnego rodzaju depozytowe mechanizmy zabezpieczające, które redukują ryzyko potencjalnego niedotrzymania warunków kontraktu. Wiele giełd energii oferuje również instrumenty pochodne o czysto finansowym charakterze, stosowane do zarządzania ryzykiem długoterminowym portfela zamówień energii.

1.2.3.2. Struktura rynku giełdowego – Towarowa Giełda Energii SA w Warszawie

Sposoby implementacji segmentu giełdowego na hurtowym rynku energii elektrycznej zależą oczywiście od konkretnego przypadku. Mogą one różnić się między sobą detalami dotyczącymi warunków funkcjonowania poszczególnych instrumentów, działania mechanizmu aukcyjnego czy też szczegółami w organizacji procesu giełdowego. Istnieje jednak pewien wspólny schemat konstrukcji rozwiązań giełdowych na rynku energii elektrycznej, stosowany w zdecydowanej większości przypadków. W miarę standardowy charakter dotyczy przede wszystkim tego, co nas będzie interesować najbardziej, czyli mechanizmu ustalania równowagi popytowo-cenowej.

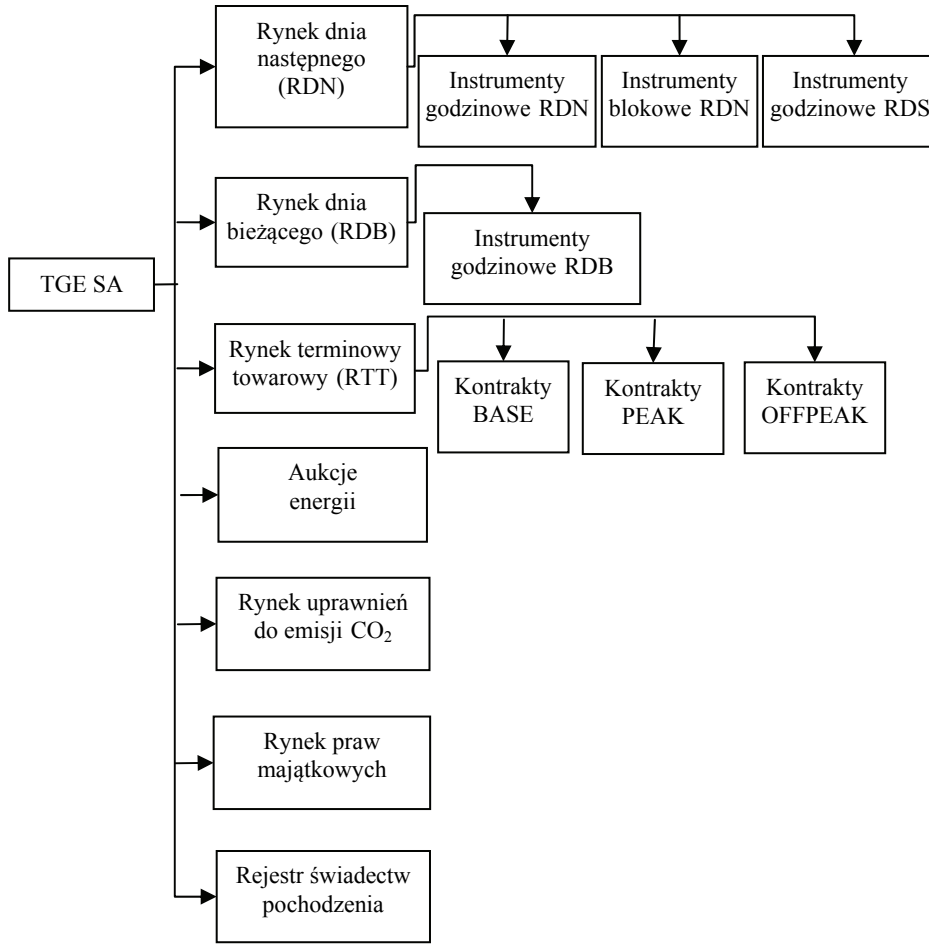
W naszej pracy, aby w sposób bardziej określony przedstawić zasadę działania segmentu giełdowego, skupimy się, ze zrozumiałych względów, przede wszystkim na mechanizmach funkcjonowania Towarowej Giełdy Energii SA (TGE SA) w Warszawie. Stanowi ona bowiem jeden z głównych elementów polskiego rynku energii elektrycznej. Jak podano w komunikatach prasowych TGE SA, w czerwcu 2012 r. łączny udział obrotu na rynkach energii elektrycznej Towarowej Giełdy Energii SA w pokryciu całościowego godzinowego zapotrzebowania na energię w krajowym systemie elektroenergetycznym (KSE) osiągnął już niemal 85%. Przy okazji polska giełda stanowi dosyć dobry przykład rozwiniętej giełdy energii funkcjonującej w ramach konkurencyjnego rynku energii elektrycznej. Oferuje ona większość charakterystycznych produktów, a mechanizmy transakcyjne na rynkach TGE SA w dużej mierze odpowiadają standardom stosowanym w tej dziedzinie.

Towarowa Giełda Energii SA w Warszawie została zarejestrowana i rozpoczęła działalność w dniu 7 grudnia 1999 r. Pierwszy jej rynek, o charakterze *spot*, w formie rynku dnia następnego, uruchomiono już w pół roku później, tj. 30 czerwca 2000 r. TGE SA jest giełdą towarową w rozumieniu Ustawy o giełdach towarowych z dnia 26 października 2000 r. i ma licencję Komisji Nadzoru Finansowego. Od lutego 2012 r. stanowi część grupy Giełdy Papierów Wartościowych SA.

Najważniejsze elementy struktury rynków funkcjonujących na Towarowej Giełdzie Energii zostały przedstawione na rysunku 1.2.9. Rysunek ten nie obejmuje oczywiście wszystkich aspektów działalności giełdy. Wybrane zostały tylko najbardziej istotne składowe, związane bezpośrednio z handlem energią elektryczną i produktami z nią związanymi. TGE SA prowadzi wiele innych

działań, których opis w naszej pracy pominiemy. Zaliczyć można do nich np. prowadzenie działalności szkoleniowej oraz informacyjnej.

Jak więc widzimy na rysunku 1.2.9, Towarowa Giełda Energii SA prowadzi różnorodne rynki pozwalające na obrót energią elektryczną oraz produktami z nią powiązanymi. Opiszmy nieco dokładniej każdy z nich.



Rysunek 1.2.9. Najważniejsze elementy struktury rynków Towarowej Giełdy Energii SA w zakresie handlu energią elektryczną i produktami z nią związanymi
Źródło: opracowanie własne

Rynek dnia następnego (RDN) – obecnie jest podstawowym rynkiem *spot* dla energii elektrycznej w Polsce. Transakcje na nim zawierane są nie później niż w dniu poprzedzającym dobę dostawy. Prowadzone procesy handlowe dotyczą trzech rodzajów instrumentów ofertowych: godzinowych (RDN i RDS)

oraz blokowych RDN. Oferty godzinowe notowane są dla każdej godziny doby handlowej rozłącznie, tworząc w zasadzie 24 niezależne, równolegle prowadzone rynki. Instrumenty godzinowe RDN dotyczą zwykłych ofert na rynku krajowym, zaś instrumenty godzinowe RDS – specjalnych ofert łączonych dla energii i praw przesyłowych jednocześnie, w ramach procesu tzw. łączenia rynków (*Market Coupling*) w wymianie międzynarodowej na rynku RDS utworzonym przez TGE SA i współpracujące giełdy zagraniczne. Instrumenty blokowe RDN dotyczą ofert kontraktów na wybrane bloki połączonych godzin doby handlowej. TGE SA oferuje obecnie trzy rodzaje ofert blokowych: BASE (cała doba), PEAK (godziny 7–22), OFFPEAK (godziny 23–7). Notowania odbywają się w systemie kursu jednolitego (2 fixingi dla krajowych instrumentów godzinowych RDN plus trzeci dla instrumentów RDS), w systemie notowań ciągłych (z wyjątkiem instrumentów RDS) oraz w transakcjach pozasesyjnych.

Rynek dnia bieżącego (RDB) – umożliwia korygowanie pozycji kontraktowych przez uczestników rynku w dniu poprzedzającym oraz w trakcie doby realizacji dostaw energii. Obecnie na Towarowej Giełdzie Energii SA na rynku RDB notowane są instrumenty godzinowe, czyli jednorazowe kontrakty dostawy, na konkretną godzinę doby. Notowania na rynku RDB prowadzone są wyłącznie w trybie notowań ciągłych. Dla godzin nocnych i porannych sesja zamykana jest w dniu poprzedzającym dobę dostawy energii (ale później niż w przypadku RDN). Dla godzin późniejszych, począwszy od H11 (czyli 10–11), notowania prowadzone są także w dniu dostawy oraz kończą się na dwie i pół godziny wcześniej przed godziną dostawy (tzn. dla H11 notowania zamykają się ostatecznie o 7.30, dla H12 o godzinie 8.30 itd.). Jednakże dla wszystkich godzin handlowych późniejszych od H18 sesja zamykana jest ostatecznie o godzinie 14.30.

Rynek towarowy terminowy (RTT) – w odróżnieniu od poprzednio charakteryzowanych rynków funkcjonujących na Towarowej Giełdzie Energii SA, RTT umożliwia zawieranie kontraktów o dłuższym okresie trwania, pozwalających na ustalenie ceny energii elektrycznej w dłuższym horyzoncie czasowym. W obrocie na RTT pozostają kontrakty typu *forward* z fizyczną realizacją, tzn. takie, w których

> sprzedawca (wystawca kontraktu) zobowiązuje się do dostarczenia energii elektrycznej w określonym terminie w przyszłości i po określonej cenie, a nabywca (nabywca kontraktu) zobowiązuje się do nabycia energii elektrycznej w określonym terminie i po określonej cenie (TGE 2012).

Kontrakty notowane na giełdzie mają standardowy charakter. Ze względu na termin wykonania, na TGE SA obraca się kontraktami tygodniowymi, miesięcznymi, kwartalnymi i rocznymi. Dostawa może obejmować całą dobę (24 godziny) – kontrakty typu BASE7 (na wszystkie dni tygodnia) i BASE5 (na dni robocze), godziny szczytowe doby (7.00–22.00, 15 godzin) – kontrakty typu PEAK7 i PEAK5 (również na wszystkie i robocze dni tygodnia) oraz godziny

pozaszczytowe (godziny od 0.00 do 7.00 i od 22.00 do 24.00 w dni robocze oraz cała doba w dni wolne od pracy) – kontrakty typu OFFPEAK. Handel kontraktami na RTT prowadzony jest tylko w systemie notowań ciągłych.

Aukcje energii – segment ten stanowi element rynku terminowego będący rozszerzeniem systemu notowań ciągłych na kontraktach terminowych na rynku RTT. Aukcje organizowane są na zlecenie uczestników giełdy. Ich cel polega na obrocie kontraktami *forward*, przede wszystkim o dużych wolumenach energii i długich okresach trwania (kwartalnym i rocznym). Kontrakty, którymi obraca się w systemie aukcji na TGE SA, mają również charakter standaryzowany.

Rynek uprawnień emisji CO_2 (RUE) – stanowi element krajowego i europejskiego systemu handlu uprawnieniami do emisji CO_2. Notowania prowadzone są dwa razy w tygodniu (we wtorek i w czwartek) w systemie kursu jednolitego, jak również w formie notowań ciągłych oraz transakcji pozasesyjnych. Obrót dotyczy kontraktów *spot* (CO_2-2012) na dostawę uprawnień do emisji CO_2 w ramach krajowego i wspólnotowego systemu handlu uprawnieniami do emisji CO_2.

Rynek praw majątkowych (RPM) – umożliwia prowadzenie obrotu zbywalnymi świadectwami pochodzenia energii. Obecnie RPM na Towarowej Giełdzie Energii SA składa się z dwóch segmentów dla energii elektrycznej (świadectw pochodzenia dla energii elektrycznej wyprodukowanej w OZE i świadectw pochodzenia dla energii elektrycznej wyprodukowanej z wysokosprawnej kogeneracji) oraz trzeciego dla świadectw pochodzenia dla biogazu rolniczego.

Świadectwo pochodzenia stanowi dowód wyprodukowania pewnej, określonej w nim, ilości energii elektrycznej w koncesjonowanym źródle odnawialnym bądź kogeneracyjnym (produkującym w skojarzeniu energię elektryczną i cieplną). Wydawane jest ono przez prezesa Urzędu Regulacji Energetyki. Wszystkie przedsiębiorstwa energetyczne, które sprzedają energię elektryczną końcowym odbiorcom, mają obowiązek uzyskania odpowiedniej (ustalanej rozporządzeniem ministra gospodarki) liczby świadectw pochodzenia i przedstawienia ich do umorzenia prezesowi URE (na podstawie art. 8, ustawy Prawo energetyczne: PE-JT-URE 2012). Rozwiązanie alternatywne polega na uiszczeniu opłaty zastępczej na konto Narodowego Funduszu Ochrony Środowiska i Gospodarki Wodnej. Prawa majątkowe wynikające ze świadectw pochodzenia mają charakter zbywalny (art. 5a, ust. 6, ustawy Prawo energetyczne: PE-JT-URE 2012) i mogą stanowić przedmiot obrotu na rynku praw majątkowych TGE SA.

Rejestr świadectw pochodzenia (RŚP) – prawa majątkowe do świadectw pochodzenia powstają z chwilą zapisania świadectwa na koncie ewidencyjnym wytwórcy w rejestrze świadectw pochodzenia, którego prowadzenie powierzono TGE SA (TGE-RPM 2012) (na podstawie art. 5a, ust. 9 ustawy Prawo energetyczne: PE-JT-URE 2012).

Jak widzimy, giełda energii elektrycznej oferuje całą paletę wyspecjalizowanych rynków. W kolejnych punktach bieżącego podrozdziału przyjrzymy się w sposób bardziej szczegółowy sposobom funkcjonowania rynków natychmiastowych, które mają największe znaczenie dla zarządzania popytem krótkoterminowym. Najwięcej miejsca poświęcimy przy tym rynkowi dnia następnego jako podstawowemu rynkowi *spot*. Nieco miejsca zarezerwujemy również dla rynku dnia bieżącego.

1.2.3.3. Rynek dnia następnego (RDN) TGE SA

Na rynku dnia następnego TGE SA prowadzi się handel kontraktami na dostawę energii elektrycznej, przy czym transakcje kończą się nie później niż w dniu poprzedzającym dobę dostawy (dobę handlową). Konkretna sesja notowań, dotycząca określonej doby dostawy, prowadzona jest na dwa dni wcześniej oraz w dniu poprzedzającym. Jak już wspomnieliśmy w poprzednim punkcie, na rynku RDN notuje się trzy rodzaje instrumentów ofertowych podzielonych na dwie grupy: instrumenty godzinowe (krajowe RDN i międzynarodowe RDS) oraz instrumenty blokowe RDN.

Ich notowania prowadzone są codziennie w systemie kursu jednolitego (w chwili pisania tego rozdziału były to dwa fixingi dla krajowych instrumentów godzinowych RDN oraz trzeci dla instrumentów RDS) i w systemie notowań ciągłych (z wyjątkiem instrumentów RDS). Szczegóły harmonogramu sesji giełdowych dla różnych typów instrumentów znajdują się w tabelach 1.2.1–1.2.3. Transakcje instrumentami zawierane mogą być również poza okresami sesji, a następnie w określonym trybie zgłaszane i rejestrowane w systemie giełdy jako tzw. transakcje pozasesyjne.

Oferty (instrumenty) godzinowe odnoszą się do dostawy na konkretną godzinę doby handlowej, której dotyczy dana sesja giełdowa. Handel tymi instrumentami dla każdej z 24 godzin dostawy traktowany jest jako odrębny rynek. Oferty nie mogą łączyć dostawy w kilku godzinach ani być przesuwane między różnymi godzinami realizacji. Jednostką notowania instrumentów godzinowych jest jeden instrument, odpowiadający 1 MWh energii elektrycznej z dokładnością do 0,1 MWh. Zlecenia mogą być bowiem dzielone na pasma (bloki) energii o zróżnicowanych cenach oraz realizowane częściowo, przy czym każda taka częściowa transakcja dotyczy co najmniej 0,1 MWh.

Instrumenty blokowe związane są z ofertami dostaw na wybrane bloki połączonych godzin doby handlowej. Uczestnicy rynku składają zlecenia na energię, która ma zostać dostarczona w ciągu ustalonego standardowego bloku godzin (łącznie, bez rozbicia na poszczególne godziny składowe) po określonej dla tej oferty cenie. Towarowa Giełda Energii SA proponuje obecnie trzy rodzaje ofert blokowych: BASE – cała doba, PEAK – godziny od 7.00 do 22.00 i OFFPEAK – godziny od 23.00 do 7.00. Nominałem kontraktów blokowych na rynku dnia

następnego jest ilość energii elektrycznej (w MWh) wyrażona jako iloczyn 1MW oraz liczby godzin w terminie realizacji danego rodzaju kontraktu.

W przypadku kontraktów blokowych typu PEAK jednostką notowania jest więc 15 MWh, ponieważ okres dostawy obejmuje 15 godzin doby handlowej. Dla kontraktów typu BASE jednostka notowania wynosi od 23 do 25 MWh, przy czym zazwyczaj są to, oczywiście, 24 MWh. Odchylenia o megawatogodzinę mniej lub więcej związane są z dniami zmiany czasu, odpowiednio, na letni i zimowy. I w końcu dla kontraktów typu OFFPEAK jednostka notowania wynosi od 8 do 10 MWh, które to widełki ponownie wynikają z różnej liczby godzin wykonania w dobach zmiany czasu.

Przyjrzyjmy się teraz sesji na rynku dnia następnego TGE SA. Harmonogramy notowań wyglądają nieco odmiennie dla poszczególnych rodzajów instrumentów.

Dla **instrumentu godzinowego RDN** notowania odbywają się zarówno w systemie kursu jednolitego (2 fixingi), jak i w systemie notowań ciągłych. Harmonogram sesji znajduje się w tabeli 1.2.1. Jak widzimy, na jej strukturę składają się przeplatane okresy notowań ciągłych i aukcji, na których ustalany jest kurs jednolity. Transakcje na dobę handlową k rozpoczynają się już w dniu $k - 2$, od długiej fazy notowań ciągłych prowadzonej od godziny 7.15 do 14.30. Następnie, w dniu $k - 1$, mamy dwie aukcje, na których ustalany jest kurs jednolity, prowadzone, odpowiednio, w godzinach 7.15–8.00 oraz 10.15–10.30.

Tabela 1.2.1. Harmonogram notowań instrumentów godzinowych RDN na TGE SA

Godziny	Faza notowań
1	2
Do 18.30 na 3 dni przed dniem dostawy	AKTUALIZACJA ZABEZPIECZEŃ Wprowadzenie aktualnych zabezpieczeń
Od 7.15 na 2 dni przed dniem dostawy do 14.30 na 2 dni przed dniem dostawy	NOTOWANIA CIĄGŁE Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń
Do 18.30 na 2 dni przed dniem dostawy	AKTUALIZACJA ZABEZPIECZEŃ Wprowadzenie aktualnych zabezpieczeń
Od 7.15 na 1 dzień przed dniem dostawy do 8.00 na 1 dzień przed dniem dostawy	FAZA PRZED NOTOWANIAMI W SYSTEMIE KURSU JEDNOLITEGO Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń
8:00 na 1 dzień przed dniem dostawy	OKREŚLENIE PIERWSZEGO KURSU JEDNOLITEGO Określenie kursów dla wszystkich godzin dnia dostawy i podanie wyników notowań na niepublicznej stronie internetowej
Od 8.01 na 1 dzień przed dniem dostawy do 10.15 na 1 dzień przed dniem dostawy	NOTOWANIA CIĄGŁE Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń
Od 10.15 na 1 dzień przed dniem dostawy do 10.30 na 1 dzień przed dniem dostawy	FAZA PRZED NOTOWANIAMI W SYSTEMIE KURSU JEDNOLITEGO Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń

Tabela 1.2.1 (cd.)

1	2
10.30 na 1 dzień przed dniem dostawy	OKREŚLENIE DRUGIEGO KURSU JEDNOLITEGO Określenie kursów dla wszystkich godzin dnia dostawy i podanie wyników notowań na niepublicznej stronie internetowej
Od 10.31 na 1 dzień przed dniem dostawy do 13.30 na 1 dzień przed dniem dostawy	NOTOWANIA CIĄGŁE Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń
Do 13.50 na 1 dzień przed dniem dostawy albo zgodnie z komunikatem	AKTUALIZACJA GRAFIKÓW PRACY PRZEZ CZŁONKÓW GIEŁDY
Do 14.30 na 1 dzień przed dniem dostawy	ZGŁASZANIE TRANSAKCJI HANDLOWYCH DO OSP
Do 17.00 na 1 dzień przed dniem dostawy	OPUBLIKOWANIE WYNIKÓW NOTOWAŃ NA PUBLICZNEJ STRONIE INTERNETOWEJ

Źródło: Towarowa Giełda Energii SA, *Szczegółowe zasady obrotu i rozliczeń dla energii elektrycznej na rynku dnia następnego*, z dnia 29 maja 2012 r., weszły w życie z dniem 11 czerwca 2012 r.

Pomiędzy obydwoma fixingami, od 8.01 do 10.15, mamy drugą, krótką fazę notowań ciągłych. Handel na rynku RDN kończy się ostatnim, dłuższym okresem notowań ciągłych odbywającym się w godzinach 10.31–13.30. Po tej godzinie nie ma już możliwości zawierania transakcji na rynku RDN (transakcje pozasesyjne muszą być zgłaszane i potwierdzane do godziny 13.15). W późniejszym okresie uczestnicy rynku aktualizują już tylko swoje grafiki zgłaszane operatorowi systemu przesyłowego, uwzględniając zawarte transakcje giełdowe, po czym następuje przekazanie informacji o wynikach sesji do OSP oraz ich ogłoszenie na publicznie dostępnej stronie internetowej TGE SA.

Jak więc widzimy, uczestnicy rynku mają spory zestaw możliwości wyboru czasu i formy składania zleceń. Pamiętajmy, że dochodzi jeszcze do tego możliwość prowadzenia przez giełdę transakcji pozasesyjnych. W praktyce jednak na rynku dnia następnego dają się zauważyć dwie tendencje. Po pierwsze, uczestnicy rynku preferują raczej handel w dniu $k-1$. Wynika to z faktu, że w procesie składania ofert mogą wtedy wykorzystać najnowsze informacje, które mają wpływ na określenie wielkości wolumenu oferty oraz proponowanych przez nich cen, np. dotyczące spodziewanego zapotrzebowania na energię przez odbiorców, ograniczeń technicznych w systemie itp. Po drugie, obserwując wolumen transakcji na rynku RDN, można wyraźnie dostrzec, że uczestnicy preferują handel na aukcjach po cenach jednolitych. Ilości energii, którymi obraca się w transakcjach zawieranych w trakcie sesji notowań ciągłych, stanowią zazwyczaj kilka procent całości energii elektrycznej sprzedawanej na

rynku dnia następnego Towarowej Giełdy Energii SA (biorąc pod uwagę instrumenty godzinowe).

Zauważmy jeszcze w harmonogramie sesji rynku dnia następnego TGE SA powtarzającą się pozycję: „aktualizacja zabezpieczeń". Każde zlecenie kupna sprawdzane jest pod kątem jego wartości oraz innych zleceń kupna złożonych wcześniej przez składającego je uczestnika, by zweryfikować, czy posiada on na swoim koncie transakcyjnym dostateczne środki na zabezpieczenie transakcji. Wysokość limitu transakcyjnego musi zostać dostosowana przez uczestnika rynku w odpowiednim momencie, przed rozpoczęciem danej fazy notowań.

Instrumenty godzinowe RDS dotyczą specjalnych ofert łączonych, jednocześnie dla energii i praw przesyłowych, w ramach procesu tzw. łączenia rynków (*Market Coupling*) w wymianie międzynarodowej na rynku RDS utworzonym przez TGE SA i współpracujące giełdy zagraniczne. Łączenie rynków stanowi istotną inicjatywę, której cel polega na doprowadzeniu w perspektywie kilku kolejnych lat do znacznego zwiększenia integracji rynków energii elektrycznej w Unii Europejskiej (Brandt 2012). Szersza analiza tego obszernego zagadnienia pozostaje jednak poza zakresem tematycznym naszej pracy.

W chwili obecnej instrumenty RDS związane są z wymianą handlową na połączeniu kablowym Polski ze Szwecją (Swe-Pol Link), rozpoczętą 15 grudnia 2010 r. Jak do tej pory, rynek ten wykorzystywany jest przede wszystkim do importu energii elektrycznej przez nabywców na TGE SA. Dla porównania, z analizy przepływów energii w trakcie mniej więcej półtorarocznego okresu funkcjonowania rynku RDS (do czerwca 2012 r.) wynika, że sumaryczna wielkość eksportu z Polski do Szwecji wyniosła 378 GWh, zaś import do Polski 2641 GWh (Brandt 2012).

Tym niemniej, jak widzimy, w formie wymiany międzynarodowej w ramach łączenia rynków już obecnie obraca się sporymi ilościami energii. Rynek ten stanowi dzisiaj znaczną część rynku RDN na Towarowej Giełdzie Energii SA. Udział transakcji RDS w całości obrotu energią na giełdowym rynku *spot* w niektórych dobach handlowych sięga nawet poziomu dwudziestu kilku – trzydziestu procent. Niestety, pewną barierą rozwojową tej inicjatywy są stosunkowo ograniczone możliwości przesyłowe na kablu Swe-Pol Link.

Tabela 1.2.2. Harmonogram notowań instrumentów godzinowych RDS na TGE SA

Godziny	Faza notowań
1	2
Do 18.30 na 2 dni przed dniem dostawy	AKTUALIZACJA ZABEZPIECZEŃ Wprowadzenie aktualnych zabezpieczeń
Od 10.35 na 1 dzień przed dniem dostawy do 11.30 na 1 dzień przed dniem dostawy	FAZA PRZED NOTOWANIAMI W SYSTEMIE KURSU JEDNOLITEGO Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń

Tabela 1.2.2 (cd.)

1	2
W terminie podanym w komunikacie na 1 dzień przed dniem dostawy	OKREŚLENIE CENY ROZLICZENIOWEJ Określenie kursów dla wszystkich godzin dnia dostawy i podanie wstępnych (nietransakcyjnych) oraz ostatecznych (transakcyjnych) wyników notowań na niepublicznej stronie internetowej
Do 13.50 na 1 dzień przed dniem dostawy albo zgodnie z komunikatem	AKTUALIZACJA GRAFIKÓW PRACY PRZEZ CZŁONKÓW GIEŁDY
Do 14.30 na 1 dzień przed dniem dostawy	ZGŁASZANIE TRANSAKCJI HANDLOWYCH DO OSP
Do 17.00 na 1 dzień przed dniem dostawy	OPUBLIKOWANIE WYNIKÓW NOTOWAŃ NA PUBLICZNEJ STRONIE INTERNETOWEJ

Źródło: Towarowa Giełda Energii SA, *Szczegółowe zasady obrotu i rozliczeń dla energii elektrycznej na rynku dnia następnego*, z dnia 29 maja 2012 r., weszły w życie z dniem 11 czerwca 2012 r.

Jak już wspomnieliśmy, instrumenty RDS mają charakter godzinowy. W przeciwieństwie jednak do krajowych instrumentów godzinowych RDN, obrót nimi odbywa się wyłącznie na jednej aukcji, w systemie kursu jednolitego. Zlecenia na dobę handlową k składane są do godziny 11.30 dnia poprzedniego. Szczegółowy harmonogram notowań tego instrumentu przedstawiony został w tabeli 1.2.2.

Notowania wszystkich trzech rodzajów **instrumentów blokowych RDN** na Towarowej Giełdzie Energii SA prowadzone są wyłącznie w systemie notowań ciągłych. Harmonogram sesji dla tego typu ofert przedstawiony został w tabeli 1.2.3. Jak widzimy, dla dostaw energii elektrycznej realizowanych w dniu k notowania ofert blokowych odbywają się w dwóch okresach notowań ciągłych: pierwszy w dniu $k-2$, w godzinach 7.15–14.30, oraz drugi w dniu $k-1$, od 7.15 do 13.00.

Tabela 1.2.3. Harmonogram notowań instrumentów blokowych RDN na TGE SA

Godziny	Faza notowań
1	2
Do 18.30 na 3 dni przed dniem dostawy	AKTUALIZACJA ZABEZPIECZEŃ Wprowadzenie aktualnych zabezpieczeń
Od 7.15 na 2 dni przed dniem dostawy do 14.30 na 2 dni przed dniem dostawy	NOTOWANIA CIĄGŁE Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń
Do 18.30 na 2 dni przed dniem dostawy	AKTUALIZACJA ZABEZPIECZEŃ Wprowadzenie aktualnych zabezpieczeń

1	2
Od 7.15 na 1 dzień przed dniem dostawy do 13.00 na 1 dzień przed dniem dostawy	NOTOWANIA CIĄGŁE Przyjmowanie zleceń; zlecenia można usuwać i modyfikować; zlecenia są sprawdzane ze względu na stan zabezpieczeń
Do 13.50 na 1 dzień przed dniem dostawy albo zgodnie z komunikatem	AKTUALIZACJA GRAFIKÓW PRACY PRZEZ CZŁONKÓW GIEŁDY
Do 14.30 na 1 dzień przed dniem dostawy	ZGŁASZANIE TRANSAKCJI HANDLOWYCH DO OSP
Do 17.00 na 1 dzień przed dniem dostawy	OPUBLIKOWANIE WYNIKÓW NOTOWAŃ NA PUBLICZNEJ STRONIE INTERNETOWEJ

Źródło: Towarowa Giełda Energii SA, *Szczegółowe zasady obrotu i rozliczeń dla energii elektrycznej na rynku dnia następnego*, z dnia 29 maja 2012 r., weszły w życie z dniem 11 czerwca 2012 r.

Zlecenia transakcji poszczególnymi instrumentami mogą więc uczestniczyć w różnych fazach rynku giełdowego RDN Towarowej Giełdy Energii SA. Niektóre z nich biorą udział wyłącznie w notowaniach w systemie jednolitym, inne w systemie ciągłym, a jeszcze inne mogą uczestniczyć w obu fazach obrotu. W wielu przypadkach istnieje w tym zakresie pewna dowolność, zwłaszcza że zlecenia mogą być realizowane w części. Zleceniodawca musi więc z góry określić warunki realizacji swoich zleceń oraz termin ich ważności, z punktu widzenia udziału w poszczególnych fazach notowań na parkiecie giełdowym. Pod tym względem, zgodnie ze *Szczegółowymi zasadami obrotu i rozliczeń dla energii elektrycznej na rynku dnia następnego* (TGE-RDN 2012), możemy wyróżnić następujące typy zleceń:

1. Zlecenie dzienne (*rest of day*) – ważne jest w dniu przekazania na giełdę; może zostać złożone w dowolnej fazie sesji. Zlecenia tego typu mogą uczestniczyć zarówno w fazie notowań ciągłych, jak i w fixingach kursu jednolitego. Część zlecenia, niezrealizowana w fazie kursu jednolitego, przechodzi do fazy notowań ciągłych; podobnie w odwrotnej sytuacji, część zlecenia, która nie została zrealizowana w fazie notowań ciągłych, przechodzi następnie do fazy kursu jednolitego.

2. Zlecenie ważne do końca okresu notowań (*good until expiry*) – ważne jest do końca notowania instrumentu. Zasady jego funkcjonowania są podobne do poprzedniego przypadku; może więc ono być składane w dowolnej fazie sesji; jego niezrealizowana część może swobodnie przechodzić między fazami notowań ciągłych i kursu jednolitego. Nowy element stanowi możliwość przejścia niezrealizowanej części zlecenia na koleją sesję notowania instrumentu.

3. Zlecenie do dnia (*good until date*) – ważne jest do daty określonej na etapie składania zlecenia. Charakter jego funkcjonowania na rynku podobny jest do poprzedniego przypadku, czyli może być ono składane w dowolnej fazie sesji, jego niezrealizowana część może swobodnie przechodzić zarówno między

fazami notowań ciągłych i kursu jednolitego, jak i między kolejnymi sesjami notowania instrumentu. W tym ostatnim przypadku występuje jednak pewne ograniczenie – zlecenie bierze udział w notowaniu do dnia, w którym upływa zawarty w nim termin ważności.

4. Zlecenie czasowe (*timed order*) – ważne jest wyłącznie w dniu przekazania na giełdę, do czasu ważności określonego na etapie składania zlecenia. Może uczestniczyć wyłącznie w fazie notowań ciągłych.

5. Zlecenie tylko na aukcje (*call auction*) – ważne jest w dniu złożenia na giełdę i może uczestniczyć tylko w danej konkretnej aukcji kursu jednolitego (tylko w jednej); jego niezrealizowana część jest usuwana.

6. Zlecenie typu zrealizuj i anuluj (*fill and kill*) – uczestniczy tylko w systemie notowań ciągłych. Pozostaje ważne do chwili zawarcia pierwszej transakcji (lub kilku pierwszych transakcji, jeżeli jest realizowane w wielu transakcjach jednocześnie), przy czym niezrealizowana część zlecenia jest anulowana – zlecenie może więc być realizowane w całości, częściowo lub może nie zostać zrealizowane w ogóle. Zlecenie tego rodzaju można złożyć bez podania limitu ceny.

7. Zlecenie typu zrealizuj lub anuluj (*fill or kill*) – podobnie jak poprzednie uczestniczy tylko w systemie notowań ciągłych oraz ważne jest do chwili zawarcia pierwszej transakcji (lub kilku pierwszych transakcji, jeżeli jest ono realizowane w kilku transakcjach jednocześnie). W tym jednak przypadku zlecenie musi być zrealizowane w całości albo nie zostanie zrealizowane w ogóle; jeżeli układ innych zleceń nie pozwala na realizację zlecenia w całości, jest ono anulowane. Zlecenia dwóch ostatnich rodzajów („zrealizuj i anuluj" oraz „zrealizuj lub anuluj") nie są ujmowane w tabeli zleceń sesji giełdowej; po ich złożeniu albo następuje zawarcie transakcji, albo są one usuwane.

Uczestnik rynku w swoim zleceniu może również określić dodatkowy warunek jego aktywacji – funkcja „*Stop Loss*". Określany jest w niej poziom kursu energii elektrycznej na rynku dnia następnego, przy którym zlecenie stanie się aktywne i pojawi się na rynku.

Zgodnie z regulaminem rynku RDN (TGE-RDN 2012), jego uczestnik „składając zlecenie w systemie kursu jednolitego określa limit ceny i ilość energii będącej przedmiotem zlecenia poprzez podanie punktów granicznych wyznaczających krzywą popytu lub podaży dla jego zlecenia". Oznacza to, że może on podzielić swoje zlecenie na pasma (bloki) o różnych wolumenach energii elektrycznej, dla których może wyznaczyć różne ceny ofertowe. Dzięki temu uczestnicy rynku ograniczają ryzyko niezrealizowania zlecenia przy powszechnie obowiązującej na giełdach energii, w tym również na Towarowej Giełdzie Energii SA, zasadzie cen krańcowych i akceptacji ofert w aukcjach w systemie kursu jednolitego.

Nabywcy mają prawo podzielić zlecenie na bloki o zróżnicowanych cenach, na część z nich wystawiając niższe, „bezpieczniejsze" ceny, tak by bloki te

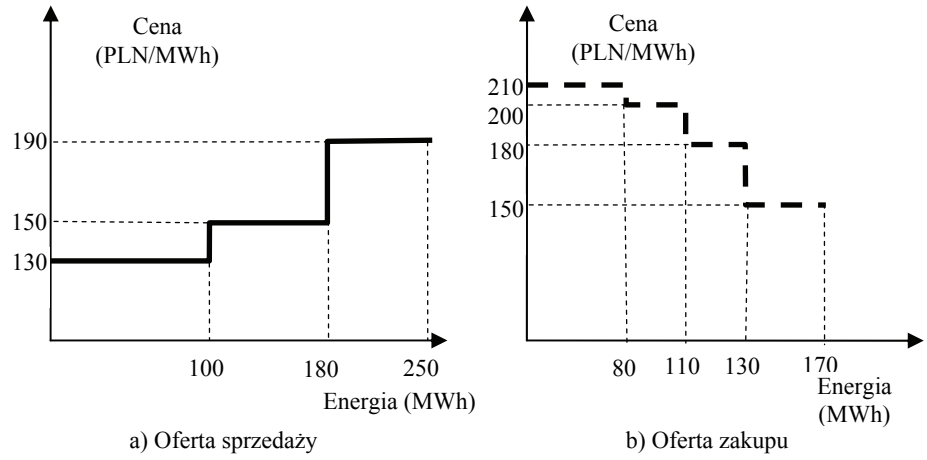
niemal na pewno weszły do realizacji, na inne natomiast wyznaczając ceny wyższe, bardziej agresywne, aby nie dopuścić do ustalenia się równowagi rynkowej na zbyt niskim poziomie cenowym, co rodzi jednak ryzyko, że ta część ich oferty nie zostanie zaakceptowana. Analogicznie nabywcy wydzielić mogą pasma bezpieczne o wyższych cenach ofertowych oraz bardziej ryzykowne, w których za pomocą niższych cen ofertowych próbują obniżyć poziom ceny równowagi rynku.

Bloki zlecenia sprzedaży ustawiane są w kolejności rosnącej, od najniższej ceny do najwyższej, zaś zakupu, odwrotnie, w kolejności malejącej, od ceny najwyższej do najniższej. W obydwu przypadkach tworzy się więc pewna krzywa schodkowa (łamana), przy czym:

– dla oferty sprzedaży jest to krzywa schodkowa rosnąca, która odzwierciedla profil podaży składanego zlecenia,

– dla oferty zakupu jest to krzywa schodkowa malejąca, która odzwierciedla profil popytu zlecenia.

Przykładowe zamówienia sprzedaży i zakupu oraz odpowiednie krzywe podaży i popytu (określane również czasem profilami zlecenia) zostały przedstawione na rysunku 1.2.10. Punkty wyznaczające wierzchołki łamanej profilu zamówienia określane są przez zleceniodawców w ofertach aukcyjnych w systemie kursu jednolitego na rynku RDN warszawskiej giełdy energii jako wspomniane punkty graniczne limitu ceny i ilości energii będących przedmiotem oferty.



Rysunek 1.2.10. Przykładowe krzywe schodkowe: podaży dla zlecenia sprzedaży i popytu dla zlecenia zakupu

Źródło: opracowanie własne

Kolejną kwestię związaną z działaniem rynku dnia następnego na Towarowej Giełdzie Energii SA stanowi sposób prowadzenia procesu handlowego na parkiecie. I znów musimy rozróżnić tutaj dwa przypadki: zasady ustalania kursów (cen) równowagi energii elektrycznej na aukcjach w systemie kursu jednolitego oraz sposób ustalania kursu i realizacja zleceń w systemie notowań ciągłych.

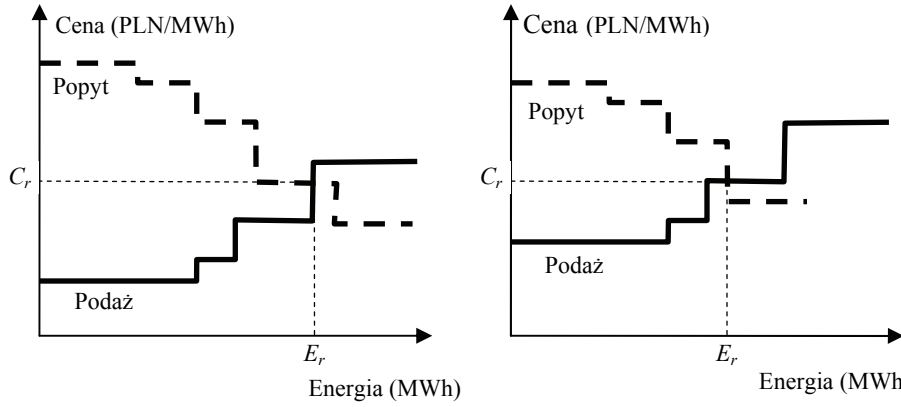
Przy ustaleniu kursu jednolitego na rynku RDN dla danej godziny doby handlowej scala się wszystkie bloki ofert złożonych dla tej godziny w jedną dużą krzywą podaży i krzywą popytu, zgodnie z regułami określonymi w paragrafie 26 *Szczegółowych zasadach obrotu i rozliczeń dla energii elektrycznej na rynku dnia następnego* (TGE-RDN 2012):

– krzywa podaży wyznaczana jest pomiędzy punktem dla teoretycznego zlecenia sprzedaży z limitem ceny równym cenie minimalnej i z wolumenem równym zero, punktami wynikającymi z ceny i wolumenu przyjętych zleceń sprzedaży oraz punktem dla teoretycznego zlecenia sprzedaży z limitem ceny równym cenie maksymalnej i z wolumenem równym skumulowanemu wolumenowi dla najwyższego limitu ceny w zleceniach sprzedaży;

– krzywa popytu wyznaczana jest pomiędzy punktem dla teoretycznego zlecenia kupna z limitem ceny równym cenie maksymalnej i z wolumenem równym zero, punktami wynikającymi z ceny i wolumenu przyjętych zleceń kupna oraz punktem dla teoretycznego zlecenia kupna z limitem ceny równym cenie minimalnej i z wolumenem równym skumulowanemu wolumenowi dla najniższego limitu ceny w zleceniach kupna.

Innymi słowy, krzywa (schodkowa) podaży zawiera ułożone niemalejąco według ceny wszystkie oferty sprzedawców (a właściwie zawarte w nich bloki – pasma – składowe), natomiast krzywa popytu – ułożone nierosnąco oferty zakupu. Poszczególne „schodki" na danym poziomie ceny scalone mogą być w wielu blokach ofertowych złożonych przez różnych odbiorców. O kolejności ich umiejscowienia decyduje czas złożenia zlecenia. Obie krzywe uzupełnione są na krańcach do przyjętych teoretycznych wartości minimalnej i maksymalnej ceny dla danego instrumentu. Ma to wyłącznie znaczenie techniczne dla działania algorytmu ustalania równowagi rynkowej, ponieważ równowaga ta, w zasadzie zawsze, ustala się na poziomach cenowych wynikających z ofert uczestników rynku.

Notowania w systemie kursu jednolitego na rynku dnia następnego TGE SA mają zatem charakter aukcji dwustronnej, z elastycznym popytem i podażą. Równowaga rynku ustalana jest poprzez określenie miejsca przecięcia się krzywych popytu i podaży zbudowanych dla złożonych na rynek ofert nabycia i sprzedaży. Przykłady ustalenia równowagi rynkowej w dwóch typowych przypadkach zostały przedstawione na rysunku 1.2.11. Miejsce przecięcia się krzywych popytu i podaży określa cenę rynkową (kurs) energii elektrycznej c_r dla danej godziny doby handlowej oraz wolumen sprzedanej na rynku o tej godzinie energii E_r.



a) Nadmiar ofert zakupu b) Nadmiar ofert sprzedaży

Rysunek 1.2.11. Przykłady ustalenia równowagi rynkowej w systemie kursu jednolitego na parkiecie RDN TGE SA w przypadku nadmiaru ofert zakupu i sprzedaży w punkcie równowagi
Źródło: opracowanie własne

Na giełdach energii stosuje się system cen krańcowych, tzn. cenę równowagi c_r wyznaczają wyłącznie ceny zleceń nabycia i sprzedaży w punkcie przecięcia krzywych popytu i podaży. Kurs ten ma charakter jednolity, transakcje energią dla wszystkich ofert przyjętych do realizacji na aukcji rynku RDN odbywają się po cenie (kursie) równowagi c_r (niezależnie od złożonej ceny ofertowej).

Dokładnie warunki akceptacji zleceń do realizacji na rynku RDN TGE SA sprecyzowane zostały w regulaminie giełdy (TGE-RDN 2012) w paragrafie 29:

a) w przypadku zleceń kupna w pierwszej kolejności będą realizowane zlecenia o najwyższym limicie ceny;

b) w przypadku zleceń sprzedaży w pierwszej kolejności będą realizowane zlecenia o najniższym limicie ceny;

c) zlecenia z równymi limitami ceny będą realizowane według czasu przyjęcia zlecenia do systemu informatycznego giełdy, zlecenie przyjęte wcześniej będzie zrealizowane w pierwszej kolejności;

d) zlecenia sprzedaży złożone z limitem ceny poniżej kursu energii elektrycznej będą zrealizowane w całości;

e) zlecenia kupna złożone z limitem ceny powyżej kursu energii elektrycznej będą zrealizowane w całości;

f) zlecenia kupna i sprzedaży złożone z limitem ceny równym kursowi energii elektrycznej mogą zostać zrealizowane częściowo, w całości lub mogą nie zostać zrealizowane;

g) zlecenia sprzedaży, złożone z pierwszym najniższym limitem ceny powyżej kursu energii elektrycznej, mogą zostać zrealizowane częściowo, w całości lub mogą nie zostać zrealizowane;

h) zlecenia kupna, złożone z pierwszym najwyższym limitem ceny poniżej kursu energii elektrycznej, mogą zostać zrealizowane częściowo, w całości lub mogą nie zostać zrealizowane.

Ujmując zatem rzecz ogólnie, do realizacji na rynku w danej godzinie handlowej przyjmowane są bloki ofert sprzedaży i zakupu energii elektrycznej położone na lewo od punktu równowagi (c_r, E_r), oferty położone na prawo od tego punktu są odrzucane. Innymi słowy, akceptowane są zlecenia sprzedaży o cenach nie większych niż c_r oraz zlecenia zakupu o cenach nie mniejszych niż c_r. Istnieją jednak pewne wyjątki od tej zasady, określone w punktach f–g przytoczonego paragrafu 29 regulaminu giełdy. Wynikają one z faktu, że przecięcie krzywych popytu i podaży rynku na ogół nie przypada w punkcie granicznym krzywej (wierzchołku łamanej określającym skok ceny), tylko pośrodku jej poziomego odcinka wyznaczającego poziom cenowy równowagi. W związku z tym niektóre oferty spełniające ogólne kryterium cenowe mogą nie zostać przyjęte przez rynek albo być przyjęte tylko w części.

W podobnej sytuacji w cytowanym regulaminie (TGE-RDN 2012), w paragrafie 32, określony został sposób ustalania punktu równowagi rynku w następujący sposób:

a) wyznaczane jest równanie prostej pomiędzy dwoma najbliższymi limitami cen w zleceniach sprzedaży, dla których różnica pomiędzy skumulowanym wolumenem kupna i skumulowanym wolumenem sprzedaży jest najmniejsza;

b) wyznaczane jest równanie prostej pomiędzy dwoma najbliższymi limitami cen w zleceniach kupna, dla których różnica pomiędzy skumulowanym wolumenem kupna i skumulowanym wolumenem sprzedaży jest najmniejsza;

c) wyznaczany jest punkt przecięcia krzywych określonych według zasad opisanych w lit a) i b); współrzędne tego punktu określają kurs i wolumen obrotu.

Innymi słowy, punkt równowagi rynku RDN na Towarowej Giełdzie Energii SA wyznaczany jest zawsze dokładnie przez punkt przecięcia krzywych podaży i popytu rynku. Spójrzmy ponownie na rysunek 1.2.11. Widzimy na nim dwa zdecydowanie najbardziej typowe przypadki sposobu przecięcia się obu krzywych.

W przypadku a) krzywa podaży przecina krzywą popytu we wnętrzu poziomego odcinka („schodka" łamanej), na poziomie ceny równowagi rynku c_r. W związku z tym mamy do czynienia z nadmiarem ofert zakupu w punkcie równowagi rynku i nie wszystkie z nich mogą być w pełni zrealizowane. Zgodnie z punktem c przytaczanego paragrafu 29 regulaminu (TGE-RDN 2012), o kolejności akceptacji ofert w tym przypadku decyduje czas przyjęcia poszczególnych zleceń zakupu do systemu informatycznego giełdy. Ostatnie zlecenie zakupu, najbliżej punktu równowagi, może być przyjęte (jeżeli zachodzi taka potrzeba) tylko w części potrzebnej do wyrównania skumulowanego popytu i wolumenu energii równowagi rynku, E_r.

W przypadku b) mamy sytuację odwrotną. Krzywa popytu przecina krzywą podaży we wnętrzu odcinka, na poziomie ceny równowagi rynku c_r. Tym razem więc w punkcie równowagi rynku obserwujemy nadmiar ofert sprzedaży i nie wszystkie te zlecenia będą mogły być w pełni zrealizowane. Ponownie o rea-

lizacji oferty giełdowej – tylko tym razem dotyczy to sprzedaży – decyduje czas jej zgłoszenia, zaś ostatnie zlecenie, najbliżej punktu równowagi, w razie potrzeby może zostać przyjęte tylko w części.

Należy tutaj zwrócić uwagę na fakt, że przedstawione rozwiązanie, zastosowane na TGE SA, nie ma charakteru uniwersalnego. Na wielu giełdach energii elektrycznej w przypadku wystąpienia na parkiecie nadmiaru zleceń sprzedaży lub nabycia w punkcie równowagi rynkowej przyjmuje się nieco inny sposób postępowania (Mielczarski 2000). Realizowane są wszystkie zlecenia na poziomie cenowym równowagi, dokonuje się jednak ich proporcjonalnej redukcji, w procencie odpowiadającym udziałowi wolumenu sprzedanej energii oferowanej po tej cenie do całości energii oferowanej po tej cenie.

Zauważmy jeszcze kolejny fakt. Otóż w systemie ustalania równowagi rynkowej, który został zaimplementowany na Towarowej Giełdzie Energii SA w Warszawie, dla danej godziny handlowej cena równowagi (kurs energii elektrycznej) c_r zawsze wyznaczana jest przez punkt przecięcia krzywej podażowej i popytowej. Jeśli spojrzymy jeszcze raz na rysunek 1.2.11, to zauważymy, że w przypadku wystąpienia w punkcie równowagi nadmiaru zleceń sprzedaży (punkt b) na rysunku), kurs rynkowy wyznaczany jest przez cenę ostatniej przyjętej do realizacji oferty sprzedaży. Natomiast w przypadku wystąpienia w punkcie równowagi nadmiaru zleceń zakupu (punkt a) na rysunku), cena równowagi wyznaczana jest przez cenę ostatniej przyjętej do realizacji oferty zakupu.

Ponownie należy stwierdzić, że nie jest to rozwiązanie uniwersalne. Wiele giełd energii elektrycznej przyjmuje jednolitą zasadę, zgodnie z którą cena (kurs) energii elektrycznej na rynku zawsze ustalana jest na podstawie ceny ostatniej przyjętej do wykonania oferty sprzedaży. Zastosowanie tej reguły pozwala na wyznaczenie niższej ceny giełdowej energii, przy zachowaniu zasady równowagi rynku, że żaden z jego uczestników nie sprzedaje poniżej i nie kupuje powyżej określonego przez niego limitu ceny ofertowej. Stąd też często rozwiązanie to określa się „zasadą najniższych cen".

Istnieje oczywiście wiele problemów szczegółowych związanych z ustalaniem równowagi rynkowej na aukcjach po kursie jednolitym na giełdach energii elektrycznej w rozmaitych specyficznych warunkach układu zleceń sprzedaży i zakupu. Można tu wymienić takie zagadnienia, jak koincydencja (pozioma lub pionowa) krzywych profili ofert, ogólny nadmiar zleceń któregoś rodzaju, brak punktu przecięcia profili itp. Zainteresowanych ich szczegółową analizą odsyłamy do literatury poświęconej konstrukcji hurtowych rynków energii elektrycznej (np. Mielczarski 2000) oraz do regulaminu rynku RDN na Towarowej Giełdzie Energii SA (TGE-RDN 2012).

Na koniec bieżącego podpunktu przyjrzyjmy się jeszcze krótko zasadom ustalania kursu i realizacji ofert w trakcie sesji w systemie notowań ciągłych na rynku dnia następnego Towarowej Giełdy Energii SA w Warszawie. Mechanizm

obrotu w tym systemie ma zresztą charakter dosyć standardowy i typowy dla giełd różnego rodzaju.

Handel w systemie notowań ciągłych prowadzony jest z użyciem tabeli zleceń, za pomocą systemu informatycznego giełdy. Oferty, które nie mogą zostać natychmiast zrealizowane, wstawiane są do tabeli zleceń i oczekują w niej na realizację. Sposób kojarzenia ofert dla nowego zlecenia został określony w paragrafie 33 cytowanego regulaminu (TGE-RDN 2012). Jest on dosyć jasny i przejrzysty. Przytoczmy go więc tutaj, aby zilustrować proces działania mechanizmu rynkowego w tym zakresie:

> Transakcje w systemie notowań ciągłych zawierane są po kursie równym limitowi ceny, jaki został podany w zleceniu wcześniej wprowadzonym, oczekującym w tabeli zleceń na realizację, zgodnie z następującymi zasadami:
>
> a) w pierwszej kolejności będą realizowane zlecenia o najwyższym limicie ceny w przypadku zleceń kupna i o najniższym limicie ceny w przypadku zleceń sprzedaży;
>
> b) w przypadku zleceń z równymi limitami ceny będą one realizowane według czasu przyjęcia zlecenia, zlecenie przyjęte wcześniej zostanie zrealizowane w pierwszej kolejności.

Jeżeli więc składane jest nowe zlecenie zakupu z określonym poziomem limitu ceny, w tabeli zleceń sprawdza się, czy są zlecenia sprzedaży o limicie ceny nie wyższym od jego ceny ofertowej. Jeżeli takie oferty istnieją, nowe zlecenie jest realizowane zgodnie z przytoczonymi zasadami. Jeśli nie, dodaje się je do tabeli zleceń. W przypadku zlecenia sprzedaży proces przebiega analogicznie.

1.2.3.4. Rynek dnia bieżącego (RDB) TGE SA

Operacje na podstawowym rynku transakcji *spot* Towarowej Giełdy Energii SA, czyli na rynku RDN, prowadzone są z niewielkim, ale istotnym wyprzedzeniem czasowym. Od chwili zamknięcia ostatniej możliwości modyfikacji pozycji kontraktowej (o godzinie 13.30 w dniu poprzedzającym dostawę, po zamknięciu sesji notowań ciągłych instrumentów godzinowych RDN) nie ma już możliwości zakupu ani sprzedaży energii elektrycznej na tym rynku. Między godziną 13.30 dnia poprzedzającego dostawę a (zwłaszcza) ostatnimi godzinami doby handlowej mija zatem niemal półtora dnia. W tym okresie mogą zaistnieć różnorodne zdarzenia, które mają wpływ na rzeczywiste zapotrzebowanie czy też produkcję energii, jak również na jej ceny na rynku bilansującym.

Kształtująca się w ten sposób niepewność sytuacji rynkowej, nadal określanej z wyraźnym wyprzedzeniem czasowym, powoduje powstanie istotnego ryzyka, zarówno ilościowego, jak i cenowego, związanego z uczestnictwem w rynku bilansującym. Pamiętajmy bowiem o specyfice energii elektrycznej jako towaru, związanej z koniecznością nieustannego równoważenia popytu i podaży w skali systemu krajowego i podsystemów lokalnych oraz z brakiem możliwości magazynowania energii. Sprawia ona, że rynek dostaw fizycznych jest rynkiem czasu rzeczywistego, na którym sytuacja może zmieniać się w bardzo dynamiczny sposób, niemalże z chwili na chwilę.

Tabela 1.2.4. Harmonogram notowań rynku RDB na TGE SA

Godziny	Faza notowań
Do 18.30 na 2 dni przed dniem dostawy	AKTUALIZACJA ZABEZPIECZEŃ Wprowadzenie aktualnych zabezpieczeń
Od 11.30 na 1 dzień przed dniem dostawy do 14.30 na 1 dzień przed dniem dostawy	NOTOWANIA CIĄGŁE Przyjmowanie zleceń dla wszystkich instrumentów dnia dostawy; zlecenia można usuwać i modyfikować; są one sprawdzane ze względu na stan zabezpieczeń
Do 18.30 na 1 dzień przed dniem dostawy	AKTUALIZACJA ZABEZPIECZEŃ Wprowadzenie aktualnych zabezpieczeń
Od 7.15 w dniu dostawy do 14.30 w dniu dostawy	NOTOWANIA CIĄGŁE Przyjmowanie zleceń dla instrumentów w okresach notowania podanych w tabeli poniżej; zlecenia można usuwać i modyfikować; są one sprawdzane ze względu na stan zabezpieczeń OKRES DOSTAWY NOTOWANYCH INSTRUMENTÓW* OKRES NOTOWANIA od H11 do H24 od 7.15 do 7.30 od H12 do H24 od 7.30 do 8.30 od H13 do H24 od 8.30 do 9.30 od H14 do H24 od 9.30 do 10.30 od H15 do H24 od 10.30 do 11.30 od H16 do H24 od 11.30 do 12.30 od H17 do H24 od 12.30 do 13.30 od H18 do H24 od 13.30 do 14.30
Do 15.30 w dniu dostawy	Opublikowanie wyników notowań na publicznej stronie internetowej

* HGG oznacza godzinę dostawy, np. H12 oznacza okres między godziną 11.00 a 12.00.

Źródło: Towarowa Giełda Energii SA, *Szczegółowe zasady obrotu i rozliczeń dla energii elektrycznej na rynku dnia bieżącego*, z dnia 8 grudnia 2010 r., weszły w życie z dniem 15 grudnia 2010 r.

W celu redukcji tych czynników ryzyka na rozwiniętych rynkach energii umożliwia się obrót energią elektryczną z jeszcze krótszymi horyzontami

czasowymi, nawet w trakcie doby dostawy. W ten sposób pozwala się zmniejszyć niepewność zmian popytowo-podażowych w przyszłości.

Tego rodzaju możliwości oferuje prowadzony przez Towarową Giełdę Energii SA rynek dnia bieżącego (RDB). Da się tu korygować pozycje kontraktowe nie tylko w dniu poprzedzającym, ale, dla późniejszych godzin doby handlowej, także w trakcie jej trwania. W obecnej chwili na rynku RDB TGE SA notowane są tylko instrumenty godzinowe, czyli jednorazowe kontrakty dostawy na konkretną godzinę doby handlowej.

Harmonogram notowań na rynku dnia bieżącego TGE SA przedstawiony został w tabeli 1.2.4. Jak widzimy, prowadzone są one wyłącznie w trybie notowań ciągłych. Sesja na dzień dostawy k podzielona jest na okres handlu w dniu $k - 1$ poprzedzającym dobę dostawy, w godzinach od 11.30 do 14.30 – czyli kończący się godzinę później niż na rynku dnia następnego – oraz kolejny okres handlu w dniu dostawy k. Notowania w dniu dostawy prowadzone są tylko dla godzin handlowych, począwszy od godziny H11 (czyli 10.00–11.00), przy czym dla kolejnych godzin kończą się tak, aby zachować dwuipółgodzinny margines czasowy między zakończeniem handlu a dostawą energii (tzn. dla H11 notowania zamykają się ostatecznie o 7.30, dla H12 o godzinie 8.30 itd.). Dla wszystkich jednak godzin handlowych późniejszych od H18 sesja zamykana jest ostatecznie o godzinie 14.30.

Handel instrumentami na rynku dnia bieżącego prowadzony jest w podobny sposób jak notowania ciągłe na rynku RDN TGE SA – w formie tabeli zleceń przy wykorzystaniu systemu informatycznego giełdy, zgodnie z regułami określonymi w *Szczegółowych zasadach obrotu i rozliczeń dla energii elektrycznej na rynku dnia bieżącego* (TGE-RDB 2010). Określony w tym dokumencie sposób prowadzenia sesji, akceptacji ofert w zasadzie jest identyczny jak ten opisany w końcowej części poprzedniego podpunktu.

1.2.3.5. Platforma POEE – rynek energii Giełdy Papierów Wartościowych

W polskim prawie energetycznym wprowadzano pojęcie publicznego obrotu energią jako zapewniającego „publiczny, równy dostęp do tej energii, w drodze otwartego przetargu, na rynku organizowanym przez podmiot prowadzący na terytorium Rzeczypospolitej Polskiej rynek regulowany lub na giełdach towarowych w rozumieniu ustawy z dnia 26 października 2000 r. o giełdach towarowych" (art. 49a ustawy Prawo energetyczne: PE-JT-URE 2012). Pojęcie to jest o tyle istotne, że przedsiębiorstwa energetyczne zobowiązane są (również przez art. 49a ustawy Prawo energetyczne) do tzw. obliga giełdowego, czyli do sprzedaży określonej części energii w ramach rynków obrotu publicznego.

Obecnie jedyną towarową giełdą energii w Polsce, w sensie wspomnianej ustawy o giełdach towarowych z dnia 26 października 2000 r., czyli przede

wszystkim posiadającą licencję Komisji Nadzoru Finansowego, jest TGE SA w Warszawie. Jak widzimy jednak, publiczny obrót energią prowadzony może być również przy użyciu mechanizmów nazwanych „rynkiem organizowanym przez podmiot prowadzący rynek regulowany". Pod tym pojęciem w obecnie obowiązującym prawodawstwie rozumie się twór określany wcześniej jako „internetowa platforma handlowa".

Internetowa platforma handlowa jako podmiot uprawniony do publicznego obrotu energią elektryczną pojawiła się w polskim prawie w 2010 r., kiedy to Ustawa z dnia 8 stycznia 2010 r. o zmianie ustawy Prawo energetyczne oraz o zmianie niektórych innych ustaw (Dz.U. z 2010 r., nr 21, poz. 104), wprowadziła to pojęcie w art. 49a, ust. 3 i 4 ustawy Prawo energetyczne (PE-JT-URE 2011):

> 3. Przez internetową platformę handlową rozumie się zespół środków organizacyjnych i technicznych umożliwiających obrót energią elektryczną za pomocą sieci internetowej przez bezpośrednie kojarzenie ofert kupna lub sprzedaży energii elektrycznej.
>
> 4. Prowadzący internetową platformę handlową są obowiązani zapewnić wszystkim uczestnikom obrotu energią elektryczną jednakowe warunki zawierania transakcji oraz dostęp, w tym samym czasie, do informacji rynkowych, a także jawność zasad działania oraz pobieranych opłat.

Szczegóły dotyczące działania platform internetowych unormowane są w Rozporządzeniu Ministra Gospodarki z dnia 17 września 2010 r. w sprawie określenia sposobu i trybu organizowania i przeprowadzania przetargu na sprzedaż energii elektrycznej oraz sposobu i trybu sprzedaży energii elektrycznej na internetowej platformie handlowej (Dz.U., nr 186, poz. 1246) (RO-PLAT 2010).

We wspomnianej tu nowelizacji zmieniono artykuł 49a, punkt 2, ustawy Prawo energetyczne, wymieniając internetowe platformy handlowe jako rynki uprawnione do realizacji obliga publicznego obrotu energią elektryczną. Podana jednak w niej definicja „internetowej platformy handlowej" uznana została za niedookreśloną prawnie, a przede wszystkim w umocowaniu platform handlowych jako rynków publicznego obrotu energią występowały poważne usterki prawne. Mianowicie, w liście uprawnionych miejsc sprzedaży publicznej energii wymienia się m.in.: „**na internetowej platformie handlowej na rynku regulowanym** w rozumieniu ustawy z dnia 29 lipca 2005 r. o obrocie instrumentami finansowymi" (PE-JT-URE 2011, art. 49a, ust. 2; podkreśl. – W.B.). Brak przecinka między określeniami „internetowej platformie handlowej" i „na rynku regulowanym" spowodował nieskuteczność działania znowelizowanej ustawy w stosunku do platform internetowych.

Z przedstawionych powodów w kolejnej nowelizacji ustawy Prawo energetyczne (Ustawa z dnia 19 sierpnia 2011 r. o zmianie ustawy Prawo energetyczne oraz niektórych innych ustaw (Dz.U. z 2011 r., nr 205, poz. 1208), zrezygnowa-

no z pojęcia „internetowej platformy handlowej" oraz jej definicji w punktach 3 i 4 wymienianego tutaj już kilkakrotnie artykułu 49a. Zamiast określać warunki funkcjonalne działania tego rodzaju rynków, sprecyzowano ich ramy prawne. Wprowadzono w związku z tym wspomniane już na początku bieżącego punktu pojęcie „rynku organizowanego przez podmiot prowadzący na terytorium Rzeczypospolitej Polskiej rynek regulowany", jednocześnie definiując tę konstrukcję w punkcie 44, artykułu 3 ustawy Prawo energetyczne (PE-JT-URE 2012) w następujący sposób:

> rynek organizowany przez podmiot prowadzący na terytorium Rzeczypospolitej Polskiej rynek regulowany – obrót towarami giełdowymi organizowany na podstawie przepisów ustawy z dnia 29 lipca 2005 r. o obrocie instrumentami finansowymi (Dz.U. z 2010 r., nr 211, poz. 1384 oraz z 2011 r., nr 106, poz. 622 i nr 131, poz. 763), odpowiednio, przez spółkę prowadzącą giełdę albo przez spółkę prowadzącą rynek pozagiełdowy.

Internetową platformą handlową spełniającą powyższe warunki rynku organizowanego przez podmiot prowadzący rynek regulowany jest platforma obrotu energią elektryczną – rynek energii Giełdy Papierów Wartościowych w Warszawie SA (POEE RE GPW). Rynek POEE uruchomiony został 11 grudnia 2010 r. Początkowo nawet cieszył się on większą popularnością niż TGE, wspomniane jednak wątpliwości prawne dotyczące zaliczenia platform internetowych do rynków obliga obrotu publicznego spowodowały, że prezes Urzędu Regulacji Energetyki odmówił zaliczenia POEE do tej kategorii. Na skutek (oczywiście w pewnej mierze) tych perturbacji nastąpił bujny rozwój rynków TGE, które obsługują obecnie ponad 80% zapotrzebowania na energię elektryczną. Dopiero nowelizacja ustawy Prawo energetyczne z 2011 r. zmieniła sytuację prawną, pozwalając interpretować sprzedaż energii na rynku POEE RE GPW jako wypełnienie obliga giełdowego obrotu publicznego.

Należy również nadmienić, że Giełda Papierów Wartościowych sfinalizowała w lutym 2012 r. zakup akcji Towarowej Giełdy Energii, stając się niemal wyłącznym (ponad 98% akcji) akcjonariuszem tej spółki. Zasadniczo rynki POEE i TGE mają dosyć zbliżony charakter (GPW-REK 2010) i docelowo planuje się ich integrację, jednakże w perspektywie najbliższych miesięcy, a prawdopodobnie nawet dłużej, będą one funkcjonować równolegle w obu formułach.

W skład platformy obrotu energią elektryczną na rynku energii GPW wchodzą dwa podstawowe rynki:

– rynek dobowo-godzinowy energii elektrycznej (REK GPW),
– rynek terminowy energii elektrycznej (RTE GPW).

1.2.4. Rynek bilansujący

1.2.4.1. Funkcje i struktura polskiego rynku bilansującego

Rynek bilansujący stanowi miejsce, na którym wykonuje się obsługę fizycznej realizacji transakcji finansowych energią, zawartych w pozostałych segmentach hurtowego rynku energii. To właśnie ten segment jest ostatecznym chwilowym rynkiem energii elektrycznej, na którym wykonywany jest rzeczywisty obrót wcześniej zakontraktowaną energią. Jego zadanie polega ponadto na ostatecznym zrównoważeniu wytwarzania i zapotrzebowania całego rynku energii elektrycznej oraz na rozliczeniu niezbilansowania poszczególnych uczestników. Ma więc on nieco odmienny charakter od poprzednio omawianych segmentów rynku, ponieważ jako miejsce obsługi fizycznych dostaw energii w najsilniejszy sposób łączy ze sobą elementy techniczne i handlowe. W naszej pracy, naturalnie, skupimy się przede wszystkim na tych ostatnich, sygnalizując jedynie najważniejsze kwestie o charakterze technicznym.

Podmiotem prowadzącym rynek bilansujący i jego centralnym elementem jest OSP – operator systemu przesyłowego (w Polsce to spółka Polskie Sieci Elektroenergetyczne Operator SA), ponieważ to właśnie on odpowiada za zrównoważenie systemu elektroenergetycznego kraju, a następnie na podstawie wyników działania rynku bilansującego prowadzi operacje techniczne związane z zaplanowaniem pracy systemu elektroenergetycznego w celu fizycznej realizacji zakontraktowanych dostaw energii. To z nim uczestnicy rynku energii prowadzą transakcje mające zapewnić bezpieczeństwo funkcjonowania rynku energii elektrycznej jako całości oraz umożliwić wykonanie dostaw energii. Jak już wspomnieliśmy w charakterystyce rynku bilansującego, w punkcie 1.2.1, możemy przy tym mówić o dwóch sposobach uczestnictwa w rynku bilansującym: pasywnym, związanym z rozliczaniem własnego niezbilansowania, oraz aktywnym, związanym z kształtowaniem ceny na rynku chwilowym.

Operacje związane z transakcjami na rynku bilansującym, niezależnie od szczegółów jego implementacji w danym kraju, prowadzone są z krótkim wyprzedzeniem czasowym, najwyżej na dzień przed dobą fizycznej realizacji dostaw (dobą handlową). Dla przykładu, w Polsce, z tego punktu widzenia, całość rynku bilansującego podzielona jest na dwa podstawowe podsegmenty:

– rynek bilansujący dnia następnego (RBN), na którym dokonuje się zgłoszeń pozycji kontraktowych dla transakcji energią elektryczną do operatora rynku bilansującego (OSP), w dniu poprzedzającym dobę fizycznej dostawy (dniu $n - 1$),

– rynek bilansujący dnia bieżącego (RBB), na którym uczestnicy rynku korygują swoje pozycje kontraktowe zgłoszone w dniu poprzednim, zgłaszając operatorowi rynku bilansującego transakcje z rynków dnia bieżącego.

By lepiej zrozumieć mechanizmy działania rynku bilansowego, w bieżącym podrozdziale przyjrzymy dokładniej jego strukturze, organizacji i procedurom funkcjonowania. Podobnie jak w poprzednim podrozdziale naszą dyskusję oprzemy na analizie polskiego rynku bilansującego, prowadzonego przez spółkę Polskie Sieci Elektroenergetyczne Operator SA.

Ramy prawne działania rynku bilansującego w Polsce oraz funkcji operatora systemu przesyłowego w tym zakresie określa ustawa Prawo energetyczne (Dz.U. z 2012, poz. 1059 j.t.) (PE-JT-URE 2012). W jej ramach należy zwrócić uwagę przede wszystkim na art. 9c, ust. 2, pkt 8, 9, 9a, na mocy których spółka PSE Operator, jako operator systemu przesyłowego elektroenergetycznego, odpowiedzialna jest za realizację podstawowych funkcji bilansujących na rynku energii elektrycznej:

8) zakup usług systemowych niezbędnych do prawidłowego funkcjonowania systemu elektroenergetycznego, niezawodności pracy tego systemu i utrzymania parametrów jakościowych energii elektrycznej;

9) bilansowanie systemu elektroenergetycznego, określanie i zapewnianie dostępności odpowiednich rezerw zdolności wytwórczych, przesyłowych i połączeń międzysystemowych na potrzeby równoważenia bieżącego zapotrzebowania na energię elektryczną z dostawami tej energii, zarządzanie ograniczeniami systemowymi oraz prowadzenie rozliczeń wynikających z:

a) niezbilansowania energii elektrycznej dostarczonej i pobranej z systemu elektroenergetycznego,

b) zarządzania ograniczeniami systemowymi;

9a) prowadzenie centralnego mechanizmu bilansowania handlowego.

Ponadto w ustawie Prawo energetyczne (Dz.U. z 2012, poz. 1059 j.t.) (PE-JT-URE 2012) w art. 3 zdefiniowano kilka podstawowych pojęć związanych z realizacją przedstawionych obowiązków, precyzujących podstawy działania poszczególnych elementów rynku bilansującego:

– **bilansowanie systemu** jako „działalność gospodarczą wykonywaną przez operatora systemu przesyłowego lub dystrybucyjnego w ramach świadczonych usług przesyłania lub dystrybucji, polegającą na równoważeniu zapotrzebowania na paliwa gazowe lub energię elektryczną z dostawami tych paliw lub energii" (ust. 23a),

– **zarządzanie ograniczeniami systemowymi** jako „działalność gospodarczą wykonywaną przez operatora systemu przesyłowego lub dystrybucyjnego w ramach świadczonych usług przesyłania lub dystrybucji w celu zapewnienia bezpiecznego funkcjonowania systemu gazowego albo systemu elektroenergetycznego oraz zapewnienia […] wymaganych parametrów technicznych paliw gazowych lub energii elektrycznej w przypadku wystąpienia ograniczeń technicznych w przepustowości tych systemów" (ust. 23b),

– **bilansowanie handlowe** jako „zgłaszanie operatorowi systemu przesyłowego elektroenergetycznego przez podmiot odpowiedzialny za bilansowanie

handlowe do realizacji umów sprzedaży energii elektrycznej zawartych przez użytkowników systemu i prowadzenie z nimi rozliczeń różnicy rzeczywistej ilości dostarczonej albo pobranej energii elektrycznej i wielkości określonych w tych umowach dla każdego okresu rozliczeniowego" (ust. 40),

– **centralny mechanizm bilansowania handlowego** jako „prowadzony przez operatora systemu przesyłowego, w ramach bilansowania systemu, mechanizm rozliczeń podmiotów odpowiedzialnych za bilansowanie handlowe, z tytułu niezbilansowania energii elektrycznej dostarczonej oraz pobranej przez użytkowników systemu, dla których te podmioty prowadzą bilansowanie handlowe" (ust. 41),

– **podmiot odpowiedzialny za bilansowanie handlowe** jako „osobę fizyczną lub prawną uczestniczącą w centralnym mechanizmie bilansowania handlowego na podstawie umowy z operatorem systemu przesyłowego, zajmującą się bilansowaniem handlowym użytkowników systemu" (ust. 42).

W art. 9g ust. 6 ustawy Prawo energetyczne (PE-JT-URE 2012) nałożony został również następujący obowiązek: „warunki w zakresie bilansowania systemu elektroenergetycznego […] powinny umożliwiać dokonywanie zmian grafiku handlowego w dniu jego realizacji oraz bilansowanie tego systemu także przez zmniejszenie poboru energii elektrycznej przez odbiorców niespowodowane wprowadzonymi ograniczeniami". Regulacje te spowodowały wprowadzenie rynku bilansującego dnia bieżącego (RBB), natomiast prace nad przygotowaniem funkcjonowania aktywnego popytu na rynku bilansującym trwają obecnie w spółce PSE Operator (Midera 2011).

Operator systemu przesyłowego, w myśl cytowanej ustawy Prawo energetyczne (PE-JT-URE 2012, art. 9g, ust. 1, ust. 6), ma również obowiązek opracowania odpowiedniej instrukcji ruchu i eksploatacji sieci przesyłowej, zawierającej wyodrębnioną część dotyczącą bilansowania systemu i zarządzania ograniczeniami w systemie elektroenergetycznym. Zgodnie z art. 9g, ust. 6 ustawy, powinna ona przy tym określać następujące elementy (PE-JT-URE 2012):

1) warunki, jakie muszą być spełnione w zakresie bilansowania systemu i zarządzania ograniczeniami systemowymi;
2) procedury:
   a) zgłaszania i przyjmowania przez operatora systemu przesyłowego elektroenergetycznego do realizacji umów sprzedaży oraz programów dostarczania i odbioru energii elektrycznej,
   b) zgłaszania do operatora systemu przesyłowego umów o świadczenie usług przesyłania paliw gazowych lub energii elektrycznej,
   c) bilansowania systemu, w tym sposób rozliczania kosztów jego bilansowania,
   d) zarządzania ograniczeniami systemowymi, w tym sposób rozliczania kosztów tych ograniczeń,
   e) awaryjne;
3) sposób postępowania w stanach zagrożenia bezpieczeństwa zaopatrzenia w paliwa gazowe lub energię elektryczną;

4) procedury i zakres wymiany informacji niezbędnej do bilansowania systemu i zarządzania ograniczeniami systemowymi;

5) kryteria dysponowania mocą jednostek wytwórczych energii elektrycznej, uwzględniające, w przypadku elektrowni jądrowych, wymagania w zakresie bezpieczeństwa jądrowego i ochrony radiologicznej określone przepisami ustawy z dnia 29 listopada 2000 r. – Prawo atomowe, oraz kryteria zarządzania połączeniami systemów gazowych albo systemów elektroenergetycznych;

6) sposób przekazywania użytkownikom systemu informacji o warunkach świadczenia usług przesłania energii elektrycznej oraz pracy krajowego systemu elektroenergetycznego.

Instrukcja ruchu i eksploatacji sieci przesyłowej podlega zatwierdzeniu przez prezesa Urzędu Regulacji Energetyki, stając się w ten sposób dokumentem formalnoprawnym, regulującym funkcjonowanie rynku bilansującego. Dokładnie w art. 9g, ust. 12 ustawy Prawo energetyczne (PE-JT-URE 2012) stwierdza się:

Użytkownicy systemu, w tym odbiorcy, których urządzenia, instalacje lub sieci są przyłączone do sieci operatora systemu gazowego lub systemu elektroenergetycznego, lub korzystający z usług świadczonych przez tego operatora, są obowiązani stosować się do warunków i wymagań oraz procedur postępowania i wymiany informacji określonych w instrukcji zatwierdzonej przez Prezesa Urzędu Regulacji Energetyki i ogłoszonej w Biuletynie Urzędu Regulacji Energetyki. Instrukcja ta stanowi część umowy o świadczenie usług przesyłania lub dystrybucji paliw gazowych lub energii elektrycznej lub umowy kompleksowej.

Dokumentem określającym szczegółowe regulacje rynku bilansującego jest więc wydana przez operatora systemu przesyłowego, spółkę PSE Operator SA *Instrukcja ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi* (PSE-IRiESP 2012). Zgodnie z nią, podmiotami uczestniczącymi w rynku bilansującym są uczestnicy rynku bilansującego (URB), operatorzy rynku (OR) oraz operatorzy systemu.

Uczestnikiem Rynku Bilansującego jest podmiot, który ma zawartą Umowę o świadczenie usług przesyłania z OSP […], na mocy której, w celu zapewnienia sobie zbilansowania handlowego, realizuje dostawy energii poprzez obszar Rynku Bilansującego oraz podlega rozliczeniom z tytułu działań obejmujących bilansowanie energii i zarządzanie ograniczeniami systemowymi (PSE-IRiESP 2012, p. 2.1.1.2).

Może być to podmiot fizycznie przyłączony do sieci przesyłowej lub fragmentów sieci dystrybucyjnych objętych obszarem rynku bilansującego lub podmiot nieprzyłączony, który nie ma urządzeń i instalacji fizycznie przyłączonych do sieci przesyłowej lub sieci dystrybucyjnej objętej obszarem rynku bilansującego, natomiast jest uczestnikiem rynku energii – występuje jako strona transakcji sprzedaży lub kupna energii elektrycznej, których realizacja następuje w obszarze rynku bilansującego.

Charakter uczestnictwa w rynku bilansującym zależy oczywiście od funkcji danego podmiotu. Dokładniej precyzując, zgodnie z IRiESP, uczestników rynku

bilansującego podzielić możemy na następujące rodzaje (PSE-IRiESP 2012, p. 2.1.1.4):

1. **Wytwórcy energii elektrycznej** (URB_W) – podmioty działające na rynku energii, które posiadają koncesję na wytwarzanie energii elektrycznej na terenie Polski, oraz ich jednostki wytwórcze przyłączone są do sieci elektroenergetycznej objętej obszarem rynku bilansującego.

2. **Odbiorcy energii elektrycznej** (URB_O), pośród których wyróżnić można:

a) odbiorców końcowych energii elektrycznej (URB_{OK}) – podmioty pobierające energię elektryczną wyłącznie na własny użytek; ich instalacje przyłączone są do sieci elektroenergetycznej objętej obszarem rynku bilansującego; przy czym do pojęcia użytku własnego nie zalicza się energii elektrycznej kupowanej w celu jej wykorzystania na potrzeby wytwarzania, przesyłania lub dystrybucji,

b) odbiorców sieciowych (URB_{SD}) – podmioty pełniące na obszarach sieci energetycznej poszczególnych operatorów systemów dystrybucyjnych elektroenergetycznych (OSD) funkcje sprzedawców energii elektrycznej dla odbiorców końcowych w gospodarstwach domowych, którzy nie korzystają z prawa wyboru sprzedawcy.

3. **Przedsiębiorstwa obrotu** (URB_{PO}) – podmioty będące posiadaczami koncesji na obrót energią elektryczną, które są stronami transakcji sprzedaży lub zakupu energii elektrycznej, realizowanych w obszarze rynku bilansującego.

4. **Giełdy energii** (URB_{GE}) – podmioty prowadzące giełdę towarową, w formalnym rozumieniu, czyli zgodnie z ustawą o giełdach towarowych, na której zawierane są transakcje sprzedaży i zakupu energii elektrycznej, realizowane później w obszarze rynku bilansującego.

5. **Operatorzy systemów dystrybucyjnych elektroenergetycznych**, których sieć dystrybucyjna ma bezpośrednie połączenie z siecią przesyłową (URB_{OSD}); są to przedsiębiorstwa bilansujące kupujące energię elektryczną w celu pokrywania strat powstałych w sieciach dystrybucyjnych podczas dystrybucji energii elektrycznej z wykorzystaniem tej sieci.

6. **Operator systemu przesyłowego elektroenergetycznego** (URB_{BIL}) – czyli przedsiębiorstwo bilansujące, które kupuje energię elektryczną w celu pokrywania strat powstałych w sieci przesyłowej podczas przesyłania energii elektrycznej z wykorzystaniem tej sieci.

Należy jeszcze dodać, że każdy uczestnik rynku hurtowego energii elektrycznej w Polsce zobowiązany jest spełnić określone wymagania techniczne i posiadać odpowiednie systemy informatyczne niezbędne do współpracy z operatorem rynku (systemu przesyłowego). Wśród nich wymienić należy:

– konieczność posiadania odpowiedniej klasy aparatury pomiarowej dla przepływów energii oraz systemu pomiarowo-rozliczeniowego, rejestrującego, zbierającego i przesyłającego dane pomiarowe; wymagania techniczne dla układów pomiarowych (liczników) stosowanych do rozliczeń bilansowania systemu i zarządzania ograniczeniami systemowymi oraz systemów pomiarowo-

-rozliczeniowych wykorzystywanych do wymiany danych z OSP określono w *Instrukcji ruchu i eksploatacji sieci przesyłowej* w części *Warunki korzystania, prowadzenia ruchu, eksploatacji i planowania rozwoju sieci*,

– wdrożenie systemu wymiany informacji rynku energii (WIRE); niezbędny jest on do zgłoszenia operatorowi systemu przesyłowego umów sprzedaży energii (ZUSE), ewentualnych ofert bilansujących, przesyłania na żądanie OSP pomiarów z liczników energii (DGPP); za pomocą systemu WIRE uczestnik otrzymuje informacje na temat planów koordynacyjnych, danych pomiarowych z liczników energii OSP, zatwierdzania lub odrzucania umów sprzedaży i ofert bilansujących, rozliczeń i bilansów handlowo-technicznych,

– użytkownicy aktywni, oferujący operatorowi sieci przesyłowej usługi bilansowania, muszą posiadać system operatywnej współpracy z elektrowniami (SOWE) wraz z obsługą w ruchu ciągłym, co umożliwia odbieranie poleceń OSP dotyczących wymaganej redukcji poboru mocy lub wzrostu jej wytwarzania, zgłaszanie do operatora systemu przesyłowego zdarzeń ruchowych (np. awarie, niesprawności), ograniczenie możliwości redukcji lub wzrostu obciążenia, z uwagi na bieżące uwarunkowania techniczne (takie jak remonty).

Drugą grupą podmiotów działających na rynku bilansującym są operatorzy rynku (OR).

Operatorem Rynku jest podmiot, który świadczy usługi operatorskie na rynku energii na podstawie Umowy przesyłania zawartej z OSP określającej zakres i sposób realizacji działalności operatorskiej na Rynku Bilansującym, a w przypadku gdy jego działalność operatorska dotyczy sieci dystrybucyjnej również z właściwym OSD (PSE-IRiESP 2012, p. 2.1.1.5).

Do **operatorów** rynku zalicza się następujące podmioty (PSE-IRiESP 2012; Mielczarski 2000; Szczygieł 2001):

1. Operatorzy handlowi (OH) – podmioty odpowiedzialne za dysponowanie zdolnościami wytwórczymi lub odbiorczymi tzw. jednostek grafikowych uczestników rynku bilansującego w zakresie handlowym. Dysponują oni energią elektryczną wprowadzaną na rynek lub z niego odbieraną przez reprezentowane przez siebie jednostki, formułując ich handlowe grafiki pracy, następnie przesyłając je operatorowi systemu przesyłowego lub właściwemu operatorowi systemu dystrybucyjnego. Są również stroną rozliczeń transakcji sprzedaży lub zakupu energii, które zostały zawarte za ich pośrednictwem. Każdy uczestnik rynku działający na rynku bilansującym musi wszystkim posiadanym jednostkom grafikowym zapewnić obsługę właściwych funkcji operatorskich w zakresie handlowym, przy czym może sprawować je samodzielnie albo zlecić ich realizację innym podmiotom wyspecjalizowanym w roli operatora handlowego.

2. Operatorzy handlowo-techniczni (OHT) – podmioty odpowiedzialne (na zasadzie wyłączności) za reprezentowanie jednostek grafikowych uczestników rynku bilansującego, zarówno w zakresie handlowym, jak i technicznym. Dysponują oni energią elektryczną dostarczaną na rynek lub z niego odbieraną

przez ich jednostki grafikowe oraz ich zdolnościami produkcyjnymi lub przyłączeniowymi. Tworzą zbilansowane handlowo-techniczne grafiki pracy reprezentowanych przez siebie jednostek grafikowych, weryfikując wstępnie wykonalność zawartych kontraktów, a następnie przekazują je do operatora systemu przesyłowego lub właściwego operatora systemu dystrybucyjnego. Grafiki te powinny uwzględniać charakterystyki techniczne jednostek grafikowych. Operatorzy handlowo-techniczni biorą udział w rozliczeniach na rynku bilansującym, w zakresie niezbilansowania między rzeczywistymi wolumenami dostaw energii elektrycznej a ich pozycjami kontraktowymi (wartościami wcześniej ustalonymi w grafikach). Podobnie jak w przypadku OH, każdy uczestnik rynku działający na rynku bilansującym musi wszystkim posiadanym jednostkom grafikowym zapewnić obsługę właściwych funkcji operatorskich w zakresie techniczno-handlowym, przy czym może sprawować je samodzielnie albo zlecić ich realizację innym podmiotom wyspecjalizowanym w roli operatora handlowo-technicznego.

3. Operatorzy pomiarów (OP) – podmioty odpowiedzialne za pozyskiwanie danych pomiarowych energii elektrycznej z układów pomiarowo-rozliczeniowych i przekazywanie ich do OSP lub do innego operatora prowadzącego procesy rozliczeń.

Ostatnią grupę podmiotów uczestniczących w rynku bilansującym stanowią **operatorzy systemu**. Należą do niej: operator systemu przesyłowego (OSP) i operatorzy systemów dystrybucyjnych (OSD):

1. Operatorem systemu przesyłowego jest przedsiębiorstwo elektroenergetyczne posiadające koncesję na przesyłanie energii elektrycznej, odpowiedzialne za prowadzenie ruchu sieciowego w elektroenergetycznym systemie przesyłowym, za bieżące i długookresowe bezpieczeństwo jego funkcjonowania, eksploatację, konserwację, remonty oraz niezbędną rozbudowę elementów systemu, w tym sieci przesyłowej połączeń z innymi systemami elektroenergetycznymi.

2. Operatorem systemu dystrybucyjnego jest przedsiębiorstwo elektroenergetyczne zajmujące się dystrybucją energii elektrycznej, odpowiedzialne za prowadzenie ruchu sieciowego i koordynację działania w elektroenergetycznym systemie dystrybucyjnym na określonym obszarze kraju, za bieżące i długookresowe bezpieczeństwo jego funkcjonowania, eksploatację, konserwację, remonty oraz niezbędną rozbudowę elementów systemu, w tym sieci dystrybucyjnej i połączeń z innymi systemami elektroenergetycznymi.

Obszarem rynku bilansującego jest

część systemu elektroenergetycznego, w której jest prowadzony hurtowy obrót energią elektryczną oraz w ramach której OSP równoważy bieżące zapotrzebowanie na energię elektryczną z dostawami tej energii w Krajowym Systemie Elektroenergetycznym, a także zarządza ograniczeniami systemowymi i prowadzi wynikające z tego rozliczenia, z podmiotami uczestniczącymi w Rynku Bilansującym (PSE-IRiESP 2012, p. 2.1.2.1).

Umowny punkt w sieci znajdujący się w obszarze rynku bilansującego, w którym następuje przekazanie energii elektrycznej (zarówno dostawa, jak i odbiór) pomiędzy uczestnikiem rynku bilansującego a rynkiem bilansującym, nazywany jest miejscem dostarczania energii rynku bilansującego. Miejsce dostarczania może mieć charakter pojedynczego fizycznego węzła lub grupy węzłów tworzących umowny punkt graniczny w sieci, gdzie realizowane są rzeczywiste dostawy energii z rynku i na rynek, albo też charakter wirtualny (dla pośredników w handlu energią elektryczną – np. giełd, przedsiębiorstw obrotu), gdy realizowana dostawa energii nie jest powiązana bezpośrednio z fizycznymi przepływami energii (określa się to jako tzw. punkt ponad siecią).

Wolumeny energii dostarczanej lub odbieranej w fizycznych miejscach dostarczania energii określane są na podstawie pomiarów oraz odpowiednich algorytmów wyznaczania ilości energii. Natomiast w wirtualnych miejscach dostarczania określane są one na podstawie wielkości energii wynikających z umów sprzedaży energii oraz z odpowiednich algorytmów obliczeniowych.

Podstawowy obiekt na rynku bilansującym stanowi jednostka grafikowa. Najogólniej mówiąc, jednostką grafikową nazywamy zbiór miejsc dostarczania energii rynku bilansującego, traktowanych jako jedna niepodzielna całość, względem której prowadzone są operacje na rynku. Wszelkie procesy związane z planowaniem, prowadzeniem ruchu i rozliczeń realizowanych na rynku bilansującym, a w ramach tego wyznaczaniem danych handlowych i technicznych, dotyczą poszczególnych jednostek grafikowych. W obrębie procesów prowadzonych na rynku bilansującym wyznaczane są dla nich następujące wielkości (PSE-IRiESP 2012, p. 2.1.3.14):

– planowane ilości dostaw energii, w tym deklarowana, zweryfikowana i skorygowana ilość dostaw energii,

– rzeczywiste ilości dostaw energii,

– odchylenia pomiędzy planowanymi oraz rzeczywistymi ilościami dostaw energii,

– wielkości należności i zobowiązań wynikających z odchyleń pomiędzy planowanymi i rzeczywistymi ilościami dostaw energii.

Jednostka grafikowa stanowi zatem po prostu zbiór punktów, przez które uczestnicy rynku bilansującego prowadzą swoje fizyczne transakcje – mają one być traktowane łącznie, jako jedna całość w procesach bilansujących i rozliczeniowych. Nie znaczy to jednak, że pojęcie jednostki grafikowej pokrywa się z pojęciem uczestnika rynku. Na przykład jeden uczestnik rynku może posiadać kilka jednostek grafikowych, jeżeli z jakiegoś powodu chce podzielić obszar swojej działalności na kilka obszarów bilansowania.

Jak już wspomnieliśmy, uczestnicy rynku, a co za tym idzie, ich jednostki grafikowe, mogą brać w nim udział w sposób aktywny lub pasywny.

Jednostka grafikowa bierze udział w sposób aktywny w rynku bilansującym, jeżeli uczestniczy w bilansowaniu systemu i zarządzaniu występującymi

w nim ograniczeniami systemowymi. Jednostka tego rodzaju realizuje na rynku następujące czynności (PSE-IRiESP 2012, p. 2.1.3.15, (1)):

– zgłasza operatorowi systemu przesyłowego zawarte umowy sprzedaży energii (USE),

– zgłasza operatorowi systemu przesyłowego oferty bilansujące,

– uczestniczy w obszarze rynku bilansującego, w bilansowaniu wytwarzania energii elektrycznej z jej zapotrzebowaniem,

– uczestniczy w działaniach dostosowawczych prowadzących do uwzględnienia istniejących ograniczeń systemowych,

– uczestniczy w optymalizacji rozkładu obciążeń, zgodnie z algorytmem rozdziału obciążeń podczas formułowania planów koordynacyjnych dobowych (PKD) i bieżących planów koordynacyjnych dobowych (BPKD),

– uczestniczy w rozliczeniach na rynku bilansującym w zakresie wykorzystania przyjętych ofert bilansujących i rzeczywistych odchyleń od planowanych ilości dostaw energii.

Jednostka grafikowa bierze udział w sposób pasywny w rynku bilansującym, jeżeli nie bierze udziału w bilansowaniu systemu i zarządzaniu występującymi w nim ograniczeniami systemowymi. Jednostka tego rodzaju realizuje więc z kolei na rynku bilansującym następujące czynności (PSE-IRiESP 2012, p. 2.1.3.15 (2)):

– zgłasza operatorowi systemu przesyłowego zawarte przez nią umowy sprzedaży energii (USE),

– zgłasza operatorowi systemu przesyłowego oferty bilansujące o ograniczonym zakresie przekazywanych informacji,

– uczestniczy w rozliczeniach na rynku bilansującym w zakresie rzeczywistych odchyleń od planowanych ilości dostaw energii.

Biorąc pod uwagę umiejscowienie na rynku bilansującym i fundamentalne funkcje jednostki grafikowej, możemy wyróżnić następujące ich rodzaje (PSE-IRiESP 2012, p. 2.1.3.18):

1. Jednostka grafikowa wytwórcza (JG_W) – obejmuje zbiór fizycznych miejsc dostarczania energii rynku bilansującego (FMB), w których do obszaru tego rynku przyłączone są urządzenia lub instalacje jednostek wytwórczych. Wśród nich, ze względu na szczegółowy sposób uczestnictwa w rynku, wyróżnia się jednostki grafikowe wytwórcze aktywne (JG_{Wa}), pasywne (JG_{Wp}) i rozliczeniowe (JG_{Wr}).

2. Jednostka grafikowa odbiorcza (JG_O) – obejmuje zbiór fizycznych miejsc dostarczania energii rynku bilansującego, w których do obszaru tego rynku przyłączone są urządzenia lub instalacje odbiorców energii, albo zbiór fizycznych miejsc dostarczania energii rynku bilansującego, przez które realizowana jest dostawa energii dla uczestników rynku detalicznego:

a) jednostka grafikowa odbiorcza (JG_O) jest jednostką pasywną, przyłączoną do sieci,

b) nie jest wymagane, by uczestnik rynku bilansującego był właścicielem urządzeń lub instalacji przyłączonych w miejscach dostarczania energii rynku bilansującego (MB) JG_O.

3. Jednostka grafikowa źródeł wiatrowych (JG_{ZW}) – obejmuje zbiór fizycznych miejsc dostarczania energii rynku bilansującego, w których do obszaru tego rynku przyłączone są źródła energii elektrycznej wykorzystujące energię wiatru, lub poprzez które reprezentowane są w obszarze rynku bilansującego dostawy energii pochodzące ze źródeł wykorzystujących energię wiatru. Jednostki grafikowe tego rodzaju traktowane są w wyróżniony sposób w procesach realizowanych na rynku bilansującym, takich jak zgłaszanie danych handlowych i technicznych oraz rozliczanie kosztów bilansowania systemu i kosztów ograniczeń systemowych.

4. Jednostka grafikowa wymiany międzysystemowej (JG_{WM}) obejmuje zbiór fizycznych miejsc dostarczania energii rynku bilansującego:

a) w których występują połączenia z systemami elektroenergetycznymi, gdzie ruch sieciowy jest prowadzony przez zagranicznych operatorów systemów przesyłowych lub zagranicznych operatorów systemów dystrybucyjnych,

b) poprzez które realizowane są dostawy energii w ramach wymiany międzysystemowej.

5. Jednostka grafikowa operatora systemu przesyłowego (JG_{OSP}) – obejmuje zbiór fizycznych miejsc dostarczania energii rynku bilansującego, w których do obszaru rynku bilansującego są przyłączone urządzenia lub instalacje jednostek wytwórczych lub odbiorców energii w pełni dysponowane i bezpośrednio sterowane przez OSP.

6. Jednostka grafikowa bilansująca (JG_{BI}) – obejmuje zbiór miejsc dostarczania energii rynku bilansującego, przez które jest domykany bilans energii elektrycznej w obszarze rynku bilansującego lub w danym obszarze sieci. Przez tę jednostkę operator systemu przesyłowego nabywa energię elektryczną na potrzeby pokrywania strat sieciowych, strat handlowych i potrzeb własnych, które mogą występować w sieci przesyłowej.

7. Jednostka grafikowa giełdy energii (JG_{GE}) – obejmuje zbiór wirtualnych miejsc dostarczania energii rynku bilansującego, poprzez które uczestnik rynku bilansującego prowadzący giełdę towarową realizuje w obszarze tego rynku obrót energią elektryczną „ponad siecią".

8. Jednostka grafikowa generacji zewnętrznej (JG_{GZ}) – obejmuje zbiór wirtualnych miejsc dostarczania energii rynku bilansującego reprezentujących produkcję energii przez wytwórców lub pobór energii przez odbiorców poza obszarem rynku bilansującego. Operatorami rynku dla tego typu jednostki grafikowej są OSP lub OSD, którzy pełnią w tym przypadku funkcje operatora handlowego.

1.2.4.2. Określanie pozycji kontraktowych na rynku bilansującym

Uczestnicy hurtowego rynku energii elektrycznej, którzy chcą, aby zawarte przez nich transakcje, w ramach segmentu kontraktów dwustronnych i giełdowych, zostały fizycznie zrealizowane przez operatora systemu przesyłowego na rynku bilansującym, muszą przekazać informacje handlowe na ich temat. Zgłoszenia te mają określoną formę – funkcjonują w postaci dokumentu elektronicznego w systemie informatycznym rynku. Stanowią one przy tym zobowiązanie do realizacji określonych działań handlowych i technicznych.

Zgłoszenia danych handlowych dotyczą określonych jednostek grafikowych. Przekazywane są operatorowi systemu przesyłowego przez operatorów handlowych i operatorów handlowo-technicznych dysponujących tymi jednostkami. Tak jak wspomnieliśmy wcześniej, każdy uczestnik rynku bilansującego musi albo wyznaczyć odpowiedniego operatora dysponującego tymi jednostkami, albo samemu realizować funkcje operatorskie.

Przekazanie informacji o transakcjach rynkowych przez poszczególnych uczestników przyjmuje formę zgłoszenia umów sprzedaży energii (USE), obejmującego wolumeny dostaw energii elektrycznej (oddanie i pobór energii elektrycznej) realizowanych poprzez określone jednostki grafikowe w ramach kontraktów zawartych na każdą godzinę doby handlowej. Uczestnicy rynku nie przekazują zatem operatorowi systemu przesyłowego informacji o cenach energii w zawartych umowach. Są one niepotrzebne do działania rynku bilansującego. Na nim bowiem rozlicza się jedynie różnice pomiędzy zgłoszonymi pozycjami kontraktowymi, wyznaczanymi jako łączny wolumen energii z przekazanych USE, a energią w rzeczywistości odebraną lub wytworzoną przez jednostkę grafikową w danej godzinie (odpowiednio dla jednostek odbiorczych i wytwórczych). Rozliczenie samych kontraktów dwustronnych następuje bezpośrednio między ich stronami, a kontraktów giełdowych – pomiędzy giełdą a jej uczestnikiem.

Informacje handlowe przekazywane w zgłoszeniach umów sprzedaży energii dla poszczególnych jednostek grafikowych powinny mieć charakter zbilansowany, tj. wolumeny energii dostarczanej przez jedną stronę transakcji handlowej oraz odebranej przez drugą powinny sobie odpowiadać.

Zgłoszenia umów sprzedaży energii dla poszczególnych godzin danej doby handlowej dokonywane mogą być na rynku bilansującym w ramach dwóch następujących mechanizmów rynkowych:

– rynku bilansującego dnia następnego (RBN) – zgłoszenia USE dla dnia następnego,

– w ramach rynku bilansującego dnia bieżącego (RBB) – zgłoszenia USE dla dnia bieżącego.

Należy tu zaznaczyć, że zgłoszenia umów sprzedaży energii w ramach rynku bilansującego dnia następnego są podstawowym sposobem ich dokonywania.

Zgłoszenie to jest obowiązkowe dla wszystkich jednostek grafikowych, z wyjątkiem ściśle określonych w IRiESP jednostek o specjalnym charakterze. Zgłoszenia umów sprzedaży energii dla dnia bieżącego umożliwiają modyfikację (na plus lub na minus) pozycji kontraktowej danej jednostki grafikowej, wynikającej z wcześniej zgłoszonych USE dla dnia następnego.

Tryb i harmonogram **zgłaszania umów sprzedaży energii w ramach rynku bilansującego dnia następnego** określony został w punkcie 3.1.2.1.1 *Instrukcji ruchu i eksploatacji sieci przesyłowej*. *Bilansowanie systemu i zarządzanie ograniczeniami systemowymi* (PSE-IRiESP 2012). Szczegółowy harmonogram czasowy tego procesu przedstawiony został w tabeli 1.2.5.

Tabela 1.2.5. Harmonogram zgłoszeń USE w ramach rynku bilansującego dnia następnego

Termin/okres	Działania operatorów rynku	Działania OSP
Doba n – 1 godzina 9.00		Rozpoczęcie procesu zgłaszania USE dla doby n w ramach RBN (otwarcie bramki zgłoszeniowej na RBN)
Od godziny 9.00 doby n – 1 do godziny 14.30 doby n – 1	Iteracyjnie: Przesyłanie zgłoszeń USE w ramach RBN (dokumenty ZUSE) Odbiór informacji o niezgodnościach w zgłoszeniach USE i poprawianie zgłoszeń USE	Iteracyjnie: Przyjmowanie i wstępna weryfikacja zgłoszeń USE w ramach RBN Generowanie i wysyłanie informacji o niezgodnościach w zgłoszeniach USE
Doba n – 1 godzina 14.30		Zakończenie procesu zgłaszania USE dla doby n w ramach RBN (zamknięcie bramki zgłoszeniowej na RBN)
Od godziny 14.30 doby n – 1 do godziny 15.30 doby n – 1	Odbiór informacji o przyjęciu, przyjęciu ze zmianami, odrzuceniu lub braku zgłoszenia USE w ramach RBN Odbiór informacji o przyjętych USE na RBN	Ostateczna weryfikacja zgłoszeń USE Generowanie i wysyłanie informacji o przyjęciu zgłoszenia USE, przyjęciu zgłoszenia USE ze zmianami, odrzuceniu zgłoszenia USE lub braku zgłoszenia USE Generowanie i wysyłanie informacji o przyjętych USE na RBN

Źródło: Polskie Sieci Elektroenergetyczne Operator SA, *Instrukcja ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi*, tekst jednolity obowiązujący od dnia 1 lutego 2013 r.

Zgłaszanie umów sprzedaży energii dla doby handlowej n, w ramach rynku bilansującego dnia następnego, trwa w dniu n – 1, czyli poprzedzającym dobę handlową, od godziny 9.00 do godziny 14.30. Okres ten określany jest często bramką dla zgłoszeń USE w ramach RBN. Dokumenty zgłoszeń USE otrzymane przez OSP są znakowane symbolem czasu ich dostarczenia przez operatora

danej jednostki grafikowej. Zgłoszenia umów sprzedaży energii na rynku dnia następnego dotyczą całego okresu (wszystkich 24 godzin) doby handlowej.

Operator systemu przesyłowego w okresie trwania bramki dla zgłoszeń USE w ramach RBN prowadzi weryfikację przesłanych w nich danych handlowych. W przypadku wystąpienia niezgodności operatorom rynku reprezentującym obie strony USE przesyłane są odpowiednie komunikaty.

Po zamknięciu bramki dla zgłoszeń USE w ramach RBN, tj. po godzinie 14.30 dnia poprzedzającego dobę handlową, operator systemu przesyłowego przeprowadza ostateczną weryfikację zgłoszeń umów sprzedaży energii, a następnie informuje operatorów rynku dysponujących jednostkami grafikowymi o przyjęciu, przyjęciu ze zmianami, odrzuceniu albo braku zgłoszeń USE dla tych jednostek.

Tryb i harmonogram **zgłaszania umów sprzedaży energii w ramach rynku bilansującego dnia bieżącego** określono w punkcie 3.1.2.1.2 *Instrukcji ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi* (PSE-IRiESP 2012). Szczegółowy harmonogram czasowy tego procesu przedstawiony został w tabeli 1.2.6. Charakter zgłoszeń USE w ramach rynku RBB jest nieco bardziej złożony niż w przypadku rynku dnia następnego.

Tabela 1.2.6. Harmonogram zgłoszeń USE w ramach rynku bilansującego dnia bieżącego

Termin/okres	Działania operatorów rynku	Działania OSP
Doba *n* – 1 godzina 15.30		Rozpoczęcie procesu zgłaszania USE dla doby *n* w ramach RBB (otwarcie bramki zgłoszeniowej na RBB)
Od godziny 15.30 doby n – 1 do godziny 22.00 doby *n*	Iteracyjnie: Przesyłanie zgłoszeń USE w ramach RBB (nie później niż na godzinę przed rozpoczęciem okresu zgłoszenia) Odbiór informacji o przyjęciu, przyjęciu ze zmianami lub odrzuceniu zgłoszenia USE w ramach RBB	Iteracyjnie: Przyjmowanie i weryfikacja zgłoszeń USE w ramach RBB Generowanie i wysyłanie informacji o przyjęciu zgłoszenia USE, przyjęciu zgłoszenia USE ze zmianami lub odrzuceniu zgłoszenia USE
Doba *n* godzina 22.00		Zakończenie procesu zgłaszania USE dla doby *n* w ramach RBB (zamknięcie bramki zgłoszeniowej na RBB)
Doba *n* po godzinie 22.00	Odbiór informacji o przyjęciu, przyjęciu ze zmianami lub odrzuceniu zgłoszenia USE w ramach RBB	Ostatnia iteracja weryfikacji zgłoszeń USE Generowanie i wysyłanie informacji o przyjęciu zgłoszenia USE, przyjęciu zgłoszenia USE ze zmianami lub odrzuceniu zgłoszenia USE

Źródło: Polskie Sieci Elektroenergetyczne Operator SA, *Instrukcja ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi*, tekst jednolity obowiązujący od dnia 1 lutego 2013 r.

Zgłaszanie umów sprzedaży energii dla doby handlowej *n* w ramach rynku bilansującego dnia bieżącego (bramka dla zgłoszeń USE w ramach rynku RBB) trwa od godziny 15.30 doby *n* – 1 do godziny 22.00 doby *n*, czyli danej doby handlowej. Dokumenty zgłoszeń USE otrzymane przez OSP są znakowane symbolem czasu ich dostarczenia do operatora systemu przesyłowego.

Zgłoszenie USE w ramach RBB, w przeciwieństwie do zasad obowiązujących na rynku dnia następnego, może dotyczyć tylko części doby handlowej. Dotyczyć ono mianowicie musi wybranego okresu danej doby handlowej obejmującego ciągły blok godzin, począwszy od określonej godziny *h* do ostatniej godziny tej doby. Zgłoszenie umów sprzedaży energii w ramach rynku bilansującego dnia bieżącego dotyczące okresu rozpoczynającego się o godzinie *h* danej doby handlowej powinno być dostarczone do operatora systemu przesyłowego przed rozpoczęciem godziny *h* – 1 tej doby (co najmniej godzinne wyprzedzenie czasowe).

Weryfikacja zgłoszeń umów sprzedaży energii w ramach RBB, czyli dotyczących danej doby handlowej, realizowana jest iteracyjnie, w miarę możliwości technicznych operatora systemu przesyłowego w zakresie przetwarzania dostarczonych zgłoszeń. OSP zastrzega sobie jednak, że operacja ta nie będzie realizowana częściej niż jedna iteracja procesu weryfikacji w jednej godzinie doby handlowej. W toku każdej iteracji operator systemu przesyłowego dokonuje weryfikacji zgłoszeń umów sprzedaży energii i informuje operatorów rynku o przyjęciu, przyjęciu ze zmianami albo odrzuceniu zgłoszeń USE.

Należy jednak pamiętać, że uczestnictwo jednostki grafikowej w rynku bilansującym dnia bieżącego ma charakter opcjonalny. Zgłoszenie umów sprzedaży energii w ramach RBB jest niezbędne tylko wtedy, gdy dana jednostka grafikowa zawiera transakcje zakupu lub sprzedaży energii na rynkach *spot*, dnia bieżącego, które mają zostać fizycznie zrealizowane przez operatora systemu przesyłowego.

W określonych przypadkach (prace modernizacyjne, konserwacyjne lub awarie systemów informatycznych operatora systemu przesyłowego służących do obsługi zgłoszeń umów sprzedaży energii) OSP może zawiesić przyjmowanie zgłoszeń USE na rynku bilansującym dnia bieżącego. Zawieszenie przyjmowania zgłoszeń USE w ramach RBB może dotyczyć dłuższego okresu, oznacza to zamknięcie w tym okresie bramki zgłoszeń USE w ramach RBB dla wszystkich dób handlowych.

Jak już wspomnieliśmy, zgłoszenia umów sprzedaży energii i innych dokumentów związanych z obsługą procesu określania pozycji kontraktowych uczestników rynku bilansującego przybierają postać standardowych dokumentów elektronicznych XML, tworzonych przez system wymiany informacji rynku energii (WIRE). Ich szczegółowy format i zawartość określony został w *Standardach technicznych systemu WIRE* (PSE-WIRE-STD 2010).

Typowa struktura informacyjna dokumentu ZUSE na rynku bilansującym dnia następnego obejmuje, obok części dotyczących danych identyfikacyjnych zgłaszającego, przede wszystkim wyszczególnienie wolumenów transakcji dla poszczególnych godzin doby handlowej. Dla każdej zawartej umowy na dostawę lub odbiór energii elektrycznej przez zgłaszającą jednostkę grafikową określane są nazwa i kod jednostki grafikowej dostawcy (odbiorcy) oraz dysponującego nią operatora handlowego, a następnie podawane są ilości energii kupionej/sprzedanej w każdej godzinie doby handlowej. Wielkości dostaw muszą być podane w MWh, z dokładnością do 0,001 MWh. Znaki wolumenów energii dla danego rodzaju transakcji muszą być podane w sposób ustalony dla określonego rodzaju jednostki grafikowej. Przykładowo, dla jednostki grafikowej odbiorczej energię kupowaną podaje się z plusem, zaś ewentualnie sprzedawaną – z minusem. Dla jednostek wytwórczych, odwrotnie, w przypadku sprzedaży energii jej wolumen jest dodatni, przy zakupie – ujemny (PSE-IRiESP 2012).

Na koniec zgłoszenie ZUSE zawiera sumaryczne wolumeny transakcji zgłaszającej jednostki grafikowej w poszczególnych godzinach, czyli pozycję kontraktową tej jednostki na rynku bilansującym dnia następnego. Pamiętać należy, że pozycja ta może być jeszcze korygowana poprzez zgłoszenie transakcji energią na rynku bilansującym dnia bieżącego.

W przypadku zgłoszenia dla rynku dnia bieżącego zawartość dokumentu ma podobny charakter; różni się przede wszystkim faktem, iż informacje o zawartych transakcjach dotyczą tylko części, a nie wszystkich godzin danej doby handlowej.

Zgłoszenia umów sprzedaży energii na rynku bilansującym, zarówno RBN, jak i RBB, podlegają weryfikacji przez operatora systemu przesyłowego. Weryfikacja ta ma charakter dwuetapowy (PSE-IRiESP 2012, p. 3.1.4):

– sprawdzenia poprawności zgłoszenia umowy sprzedaży energii, polegającego na formalnej kontroli poprawności danych w zgłoszeniu: np. poprawności danych jednostki grafikowej i operatora handlowego, tytułu prawnego dysponowania tą jednostką, osoby zgłaszającej umowę, poprawności danych dotyczących transakcji, jednostek, wartości, dopuszczalności transakcji (dla określonego typu jednostki grafikowej),

– sprawdzenia zgodności zgłoszenia umowy sprzedaży energii: dotyczy par jednostek grafikowych powiązanych transakcją zgłaszaną dla co najmniej jednej z nich; sprawdzeniu podlegają dwa elementy, tzn. przede wszystkim zgodność wolumenów dostaw energii elektrycznej w transakcji – wymaga się, aby ilości energii podane w obu zgłoszeniach tej samej transakcji były sobie równe; po drugie, weryfikowana jest zgodność typu transakcji, tj. wymaga się, aby znaki wolumenów dostaw energii elektrycznej w obu zgłoszeniach tej samej transakcji odpowiadały jej typowi (sprzedaż lub zakup) dla danej jednostki grafikowej.

1.2.4.3. Zgłoszenia ofert bilansujących

Charakteryzowane w poprzednim punkcie zgłoszenia umów sprzedaży energii na rynku bilansującym dnia następnego oraz (ewentualnie) dnia bieżącego służą do określenia pozycji kontraktowej uczestników rynku (a w zasadzie jednostek grafikowych) stanowiących podstawę planowania realizacji zawartych przez nie kontraktów przez operatora systemu przesyłowego. Odchylenia rzeczywistych dostaw dla jednostek grafikowych od ich pozycji kontraktowej, powstające głównie w wyniku niepewności zapotrzebowania wśród odbiorców końcowych, rozliczane są w ramach zakupu bądź sprzedaży energii elektrycznej na rynku bilansującym. Cena energii ustalana jest przy wykorzystaniu prowadzonej przez operatora systemu przesyłowego aukcji bilansującej (centralny mechanizm bilansujący).

Oferty bilansujące zgłaszane są przez biorące aktywny udział w rynku jednostki grafikowe wytwórcze. Zgłoszenia ofert bilansujących obejmują dwie części:

– część handlową – obejmującą przede wszystkim oferty wolumenów i cen zwiększenia produkcji energii elektrycznej (pasma przyrostowe) lub zmniejszenia (pasma redukcyjne),

– część techniczną – obejmującą dane techniczne służące do określenia ograniczeń systemowych w elektrowniach.

Naturalnie większą uwagę zwrócimy głównie na część handlową oferty bilansującej. Z punktu widzenia tematyki niniejszej pracy, rynek bilansujący interesuje nas przede wszystkim jako narzędzie kontrolowania niepewności zapotrzebowania na energię elektryczną na hurtowym rynku energii. Jak to niejednokrotnie wspominaliśmy, kwestie techniczne będziemy jedynie sygnalizować, w zakresie niezbędnym do zrozumienia specyfiki pracy, w tym przypadku zasad bilansowania rynku energii.

Rynek bilansujący w omawianej części ma charakter rynku dnia następnego. Proces ofertowy na wszystkie godziny określonej doby handlowej *n* prowadzony jest w ostatnim dniu poprzedzającym realizację ofert. Dokładny harmonogram zgłoszeń ofert bilansujących przedstawiony został w tabeli 1.2.7.

Tryb i harmonogram **zgłaszania ofert bilansujących w ramach rynku bilansującego** określony został w punkcie 3.1.2.2 *Instrukcji ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi* (PSE-IRiESP 2012). Zgłaszanie przez operatorów aktywnych jednostek grafikowych ofert bilansujących dla doby handlowej *n* trwa od godziny 9.00 do godziny 14.30 dnia poprzedniego. O godzinie 14.30 bramka zgłoszeniowa dla ofert bilansujących na wszystkie godziny następnej doby handlowej ulega zamknięciu.

Tabela 1.2.7. Harmonogram zgłoszeń ofert bilansujących w ramach rynku bilansującego

Termin/okres	Działania operatorów rynku	Działania OSP
Doba *n* – 1 godzina 9.00		Rozpoczęcie procesu zgłaszania ofert bilansujących dla doby *n* (otwarcie bramki zgłoszeniowej dla zgłoszeń ofert bilansujących)
Od godziny 9.00 doby *n* – 1 do godziny 14.30 doby *n* – 1	Iteracyjnie: Przesyłanie zgłoszeń ofert bilansujących – część handlowa oraz zgłoszeń ofert bilansujących – część techniczna Odbiór informacji o niezgodnościach w zgłoszeniach ofert bilansujących i poprawianie zgłoszeń ofert bilansujących	Iteracyjnie: Przyjmowanie i wstępna weryfikacja zgłoszeń ofert bilansujących – część handlowa oraz zgłoszeń ofert bilansujących – część techniczna Generowanie i wysyłanie informacji o niezgodnościach w zgłoszeniach ofert bilansujących
Doba *n* – 1 godzina 14.30		Zakończenie procesu zgłaszania ofert bilansujących dla doby *n* (zamknięcie bramki zgłoszeniowej zgłoszeń ofert bilansujących)
Od godziny 14.30 doby *n* – 1 do godziny 15.30 doby *n* – 1	Odbiór informacji o przyjęciu, odrzuceniu lub braku zgłoszenia oferty bilansującej Odbiór informacji o przyjętych ofertach bilansujących – części handlowej oraz przyjętych ofertach bilansujących – części technicznej	Ostatnia iteracja weryfikacji zgłoszeń ofert bilansujących: Generowanie i wysyłanie informacji o przyjęciu oferty bilansującej, odrzuceniu zgłoszenia oferty bilansującej lub braku zgłoszenia oferty bilansującej Generowanie i wysyłanie informacji o przyjętych ofertach bilansujących – część handlowa oraz przyjętych ofertach bilansujących – część techniczna

Źródło: Polskie Sieci Elektroenergetyczne Operator SA, *Instrukcja ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi*, tekst jednolity obowiązujący od dnia 1 lutego 2013 r.

Podobnie jak w przypadku zgłoszeń umów zakupu energii, proces obsługi zgłoszeń ofert bilansujących ma charakter wymiany komunikatów w formie dokumentów elektronicznych XML, tworzonych w ramach systemu wymiany informacji rynku energii (WIRE). Szczegółowy format i zawartość informacyjna przesyłanych dokumentów zostały określone w *Standardach technicznych systemu WIRE* (PSE-WIRE-STD 2010).

Oferta bilansująca oprócz danych identyfikacyjnych jednostki grafikowej, jej operatora oraz uczestnika rynku, do którego należy, obejmuje zgłoszenie dla każdej godziny doby handlowej najważniejszych parametrów pracy danej jednostki oraz ofert mocy oferowanej na rynku bilansującym. Moce podawane w ofercie muszą być równe stałym mocom dostaw energii na rynku bilansują-

cym przez całą godzinę zgłoszenia (PSE-IRiESP 2012), tj. odpowiadać powinny energiom godzinnym oferowanym w transakcjach bilansujących.

Szczegóły konstrukcji ofert bilansujących mogą się różnić w zależności od rodzaju jednostki grafikowej biorącej udział w bilansowaniu zapotrzebowania na energię. Ograniczymy się jedynie do charakterystyki wybranych elementów typowej oferty (dla jednostki wytwórczej aktywnej) w zakresie niezbędnym do zrozumienia reguł bilansowania hurtowego rynku energii elektrycznej w Polsce.

Oferty bilansujące dla każdej godziny doby handlowej mają charakter pasmowy, tj. oferowana w nich energia elektryczna podzielona może być na maksymalnie dziesięć, uporządkowanych w kolejności, pasm mocy godzinnej (energii), stanowiących oddzielnie przedmioty akceptacji bądź odrzucenia przez rynek. Należy zaznaczyć, że ceny ofertowe energii w kolejnych pasmach muszą być rosnące. Wymaganie to jest niezbędne, by zachować stabilne funkcjonowanie rynku, dzięki gwarancji przyrostowego charakteru handlu energią w kolejnych pasmach. Dzięki temu mechanizmowi bowiem wzrost ceny równowagi rynkowej powoduje przyjmowanie przez rynek kolejnych pasm w ofercie danej jednostki grafikowej, zaś przyjęcie przez rynek danego pasma energii w tejże ofercie gwarantuje również przyjęcie wszystkich poprzednich pasm.

Podsumowując więc naszą dyskusję nad zawartością informacyjną oferty bilansującej (wytwórczej jednostki grafikowej), podkreślmy, że obejmuje ona dla każdej godziny doby handlowej dane na temat parametrów pracy tej jednostki: jej moc dyspozycyjną, minimalną i maksymalną. Jednostki wytwórcze mogą zazwyczaj pracować z pewnym marginesem przeciążenia bądź zaniżenia generowanej mocy. Przedziały te stanowią kolejną informację określaną w ofercie. Ponadto dla każdego pasma cenowego energii (maksymalnie 10 pasm) na daną godzinę doby handlowej oferta bilansująca obejmuje następujące elementy (PSE-IRiESP 2012):

– cena ofertowa energii elektrycznej w danym paśmie (określana w zł/MWh, z dokładnością do 1 grosza/MWh), ceny te w kolejnych pasmach muszą rosnąć,

– stała, oferowana przez całą godzinę handlową (tzw. średniogodzinowa) moc netto (odpowiadająca energii godzinnej netto) w danym paśmie (wartość nieujemna, określana w MW, z dokładnością do 0,001 MW); moc netto musi być równa co do wartości ilości energii netto, jaką jednostka wytwórcza dostarczy z tego pasma na rynek bilansujący w danej godzinie handlowej, pracując ze stałą mocą brutto określoną w tym paśmie,

– stała, oferowana przez całą godzinę handlową (średniogodzinowa) moc brutto (a więc odpowiadająca energii godzinnej brutto) w danym paśmie (wartość nieujemna, określana w MW, z dokładnością do 1 MW); uwzględnia ona wykorzystanie na potrzeby własne danej jednostki wytwórczej.

Aby zrozumieć istotę działania mechanizmu centralnego bilansowania, musimy wyjaśnić pewien istotny aspekt konstrukcji ofert bilansujących. Oferta bilansująca aktywnej jednostki grafikowej nie dotyczy możliwych wzrostów

(lub redukcji) wytwarzania energii elektrycznej dla danej godziny handlowej. Pasma cenowe w ofercie bilansującej jednostki grafikowej pokrywają cały zakres zdolności wytwórczych tej jednostki. Zgodnie z regulacjami obowiązującymi na rynku bilansującym (PSE-IRiESP 2012), moc brutto oferowana w pierwszym paśmie typowej oferty (dla jednostki wytwórczej aktywnej) musi być równa mocy minimalnej pracy tej jednostki, zaś suma mocy brutto we wszystkich oferowanych pasmach musi być równa mocy maksymalnej.

W początkowym okresie działania rynku bilansującego w Polsce, do roku 2009, pasma cenowe energii w ofercie bilansującej w sposób jawny dzielone były przez oferenta na dwie wspomniane już wcześniej grupy:

– pasma redukcyjne – oferty cenowe redukcji produkcji energii elektrycznej przez daną jednostkę wytwórczą poniżej jej zadeklarowanej pozycji kontraktowej wynikającej z zawartych umów na rynku; łączna suma energii oferowanej w pasmach redukcyjnych musiała być równa pozycji kontraktowej danej jednostki,

– pasma przyrostowe – oferty cenowe zwiększenia produkcji energii elektrycznej przez daną jednostkę wytwórczą ponad jej zadeklarowaną pozycję kontraktową wynikającą z zawartych umów na rynku.

Po roku 2009, wraz ze zmianą sposobu rozliczeń za energię bilansującą na system cen krańcowych oraz pewnymi (wynikającymi z tej zmiany) modyfikacjami metod rozliczeń i ustalania równowagi popytowo-cenowej na rynku bilansującym (patrz następny punkt 1.2.4.4), konieczność jawnego rozróżniania pasm redukcyjnych i przyrostowych w ofercie bilansującej zniknęła. Pojęcia te w zasadzie znikają z procedur określających regulacje rynkowe i sposoby rozliczeń przedstawionych w *Instrukcji ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi* (PSE-IRiESP 2012). Tym niemniej w dużej mierze jest to zmiana formalna, zaś odpowiednie pasma cenowe nadal spełniają swoje funkcje ofert redukcyjnych i przyrostowych. Dla wygody więc nadal będziemy często odwoływać się do tych intuicyjnych określeń.

Jak już wspomnieliśmy, oferta bilansująca dotyczy całości zdolności produkcyjnych danej aktywnej jednostki wytwórczej, począwszy od mocy minimalnej danej jednostki w pierwszym paśmie ofertowym, poprzez kolejne pasma oferujące produkcję w danej godzinie doby handlowej, z coraz wyższą mocą. Ich kolejność wyznaczana jest przez ceny oferowanej w nich mocy (energii wytworzonej w danej godzinie). Będzie więc ona również pokrywać się z kolejnością akceptacji tych ofert przez rynek.

Jak zobaczymy, mechanizm funkcjonowania rynku bilansującego skonstruowany został tak, aby pierwsze pasma oferty bilansującej, dotyczące produkcji energii elektrycznej w zakresie mocy niezbędnej do realizacji kontraktów, zawartych przez tę jednostkę we wszystkich pozostałych segmentach rynku (pozycji kontraktowej), odpowiadały pasmom redukcyjnym w ofertach bilansu-

jących starszego typu. Kolejne pasma, definiujące oferty cenowe dostaw energii w ilościach przekraczających wolumen określony w pozycji kontraktowej tej jednostki, odpowiadają z kolei klasycznym pasmom przyrostowym.

Istota działania rynku bilansującego kryje się bowiem, w znacznej mierze, w ustaleniu cen, określanych przez aktywne jednostki grafikowe, dla poszczególnych pasm mocy w ofercie bilansującej. Jak wyjaśnimy w następnym punkcie, jeżeli równowaga rynku bilansującego ustali się na takim poziomie, że łączna moc pracy jednostki wytwórczej w danej godzinie handlowej we wszystkich przyjętych pasmach ofertowych (czyli łączna energia wyprodukowana przez tę jednostkę w tej godzinie) jest niższa od jej zobowiązań kontraktowych, to wytwórca będzie musiał pokryć powstałą różnicę operatorowi rynku bilansującego.

Pamiętać należy, rzecz jasna, że za energię oferowaną w pasmach mocy mieszczących się w zakresie pozycji kontraktowej wytwórcy (czyli dawnych pasmach redukcyjnych) otrzymuje on normalne przychody z jej sprzedaży w ramach kontraktów zawartych wcześniej na hurtowym rynku energii oraz zgłoszonych operatorowi rynku bilansującego poprzez zgłoszenie umów sprzedaży (ZUSE). Ceny w tego rodzaju pasmach należy więc interpretować jako ceny, po których wytwórcy gotowi są odsprzedać operatorowi systemu przesyłowego część swoich zobowiązań kontraktowych.

Strategie ustalania cen mocy bilansującej przez jednostkę wytwórczą w pasmach ofertowych mieszczących się w zakresie pozycji kontraktowej (redukcyjnych) powinny brać pod uwagę następujące elementy:

– interesuje nas, aby nie produkować energii, którą sprzedamy odbiorcom kontraktowym, więc cena, jaką określimy dla tego rodzaju pasm ofertowych, powinna być na tyle wysoka, aby rynek nie przyjął ich do realizacji,

– z drugiej strony, cała operacja musi się nam opłacać, więc różnica między ceną jednostkową sprzedaży, którą otrzymamy w kontrakcie, a ceną rozliczeniową energii bilansującej, którą zapłacimy operatorowi rynku, musi pokrywać koszty stałe produkcji oraz zapewniać zysk; aby to uzyskać, ceny ofertowe w tego rodzaju pasmach nie powinny więc przekraczać jednostkowych kosztów zmiennych wytwarzania.

Podsumowując, cena w pasmach mocy oferty bilansującej mieszczących się w ramach pozycji kontraktowej (redukcyjnych) powinna być zbliżona do jednostkowych kosztów zmiennych wytwarzania energii elektrycznej przez danego wytwórcę. Z punktu widzenia rynku bilansującego, jako elementu regulacji systemu elektroenergetycznego, formowanie się krzywej podażowej na tym poziomie jest właściwością pożądaną, pozwalającą na optymalny (pod względem kosztowym) dobór źródeł wytwórczych. Mogą jednak tutaj istnieć pewne uwarunkowania związane z metodą wyznaczania ceny rozliczeniowej na rynku, które sprawią, że oferenci zaniżą ceny ofertowe w pasmach redukcyj-

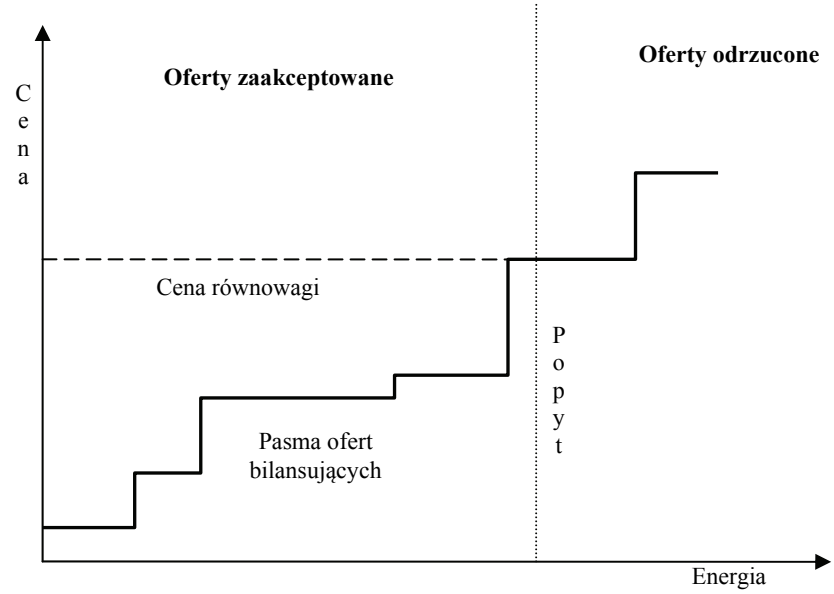
nych, by zwiększyć zyski. O problemie tym dyskutować będziemy jeszcze w następnym punkcie.

Nieco odmiennie przedstawia się sytuacja cen oferowanych przez jednostkę wytwórczą w pasmach mocy przekraczających zakres niezbędny do realizacji jej pozycji kontraktowej (przyrostowych). Jeżeli równowaga rynku bilansującego ustali się na takim poziomie, że łączna moc pracy jednostki wytwórczej w danej godzinie handlowej, we wszystkich przyjętych pasmach ofertowych, przewyższać będzie jej zobowiązania kontraktowe, to operator systemu przesyłowego kupi powstałą nadwyżkę od wytwórcy na potrzeby bilansowania rynku. Ceny pasm w części oferty przekraczającej pozycję kontraktową powinny więc być kalkulowane na poziomie co najmniej pełnych kosztów wytwarzania.

1.2.4.4. Ustalanie równowagi rynku i rozliczenia

Usługi centralnego bilansowania systemu kupowane są przez operatora rynku bilansującego na drodze aukcji jednostronnej, zgodnie ze schematem przedstawionym na rysunku 1.2.12. Pasma energetyczno-cenowe oferowane przez jednostki grafikowe, które biorą aktywny udział w bilansowaniu systemu, ustawiane są w kolejności rosnącej, według cen ofertowych aż do chwili pokrycia zapotrzebowania na energię w danej godzinie handlowej. Tworzą więc one krzywą podażową energii elektrycznej na rynku dla danej godziny doby realizacji. Popyt na rynku rzeczywistych dostaw, jak już wspomnieliśmy, charakteryzując energię elektryczną jako towar, ma charakter chwilowy oraz cechuje się bardzo niską elastycznością cenową. Punkt równowagi rynku wyznaczany jest więc przez przecięcie krzywej podażowej z prostą nieelastycznego popytu. Cena równowagi wyznaczana jest przez cenę ostatniego pasma ofertowego przyjętego do realizacji przez rynek bilansujący.

Należy przy tym nadmienić, że podobna operacja przeprowadzana jest dwukrotnie. W pierwszej iteracji kryterium porządkowania pasm energetyczno-cenowych stanowi wyłącznie cena ofertowa. Cel tej iteracji polega na wyznaczeniu ceny równowagi rynku. W kolejnej uwzględniane są ograniczenia systemowe. Pasma energetyczne niezbędne do zapewnienia fizycznej realizacji planu dostaw przesuwane są w tworzonym stosie ofert przed pasma o niższej cenie. Operacja te nie zmienia jednak ceny równowagi. Pasma energii akceptowane z powodu konieczności zachowania ograniczeń systemowych oznaczane są specjalnym znacznikiem jako wymuszone i rozliczone będą na specjalnych zasadach.



Rysunek 1.2.12. Aukcja jednostronna na rynku bilansującym
Źródło: opracowanie własne

Na rynkach bilansujących stosowane są dwie metody określania ceny rozliczeniowej (CRO) za energię bilansującą:

1. Ceny krańcowe (*marginal pricing*):

a) krańcowa cena rozliczeniowa ustalana jest jako cena ostatniej oferty wykorzystanej do zaspokojenia popytu (*Marginal Closing Price*, MCP),

b) transakcje poszczególnych aktywnych dostawców usług bilansujących oraz nieplanowane odchylenia fizycznych dostaw uczestników rynku rozliczane są według jednolitej ceny krańcowej.

2. Ceny ofertowe (*pay as bid pricing*):

a) dostawcy usług biorący udział w bilansowaniu za planowaną energię niezbędną do równoważenia rynku rozliczani są według cen pasm energetyczno-cenowych w zgłoszonych przez siebie ofertach bilansujących,

b) cena rozliczeniowa, niezbędna do rozliczeń nieplanowych odchyleń fizycznych dostaw uczestników rynku, obliczana jest jako średnia ważona (wolumenem) z poszczególnych ofert.

Najważniejszą zaletę systemu rozliczeń na rynku bilansującym na podstawie **ceny krańcowej (MCP)** stanowi doprowadzenie do sytuacji, w której oferty bilansujące jednostek wytwórczych kalkulowane są zwykle zgodnie z alternatywnymi kosztami produkcji energii. Przyjrzyjmy się odrębnie sytuacji ofert przyrostowych i redukcyjnych.

Jak stwierdziliśmy w poprzednim punkcie 1.2.4.3, ceny pasm ofertowych o mocach przekraczających wielkość pozycji kontraktowej jednostki wytwórczej (zgłoszonej operatorowi rynku bilansującego poprzez zgłoszenie umów sprzedaży) powinny być wyznaczane na poziomie co najmniej pełnych kosztów wytwarzania energii elektrycznej. W systemie MCP wytwórca, próbując maksymalizować swój zysk, nie musi ryzykować wyższych ofert cenowych, ponieważ jeśli równowaga rynkowa ustali się na poziomie ceny rozliczeniowej (CRO) przekraczającej jego ofertę cenową, i tak otrzyma płatność według wyższej ceny CRO. Dla ustalonej równowagi rynku (ceny rozliczeniowej), zysk wytwórcy z zaakceptowanych pasm bilansujących przekraczających jego pozycję ofertową nie zależy więc od ich ceny ofertowej. Może on więc w ofercie określić cenę energii w tych pasmach na minimalnym poziomie całkowitych kosztów produkcji.

Dla pasm bilansujących mieszczących się w ramach pozycji kontraktowej wytwórcy (czyli redukcyjnych) cena energii elektrycznej nie powinna być wyższa od jednostkowych kosztów zmiennych wytwarzania energii elektrycznej przez danego wytwórcę. Wynika to (jak analizowaliśmy to w poprzednim punkcie 1.2.4.3) z faktu, że ceny w tego rodzaju pasmach należy interpretować jako ceny, po których wytwórcy gotowi są odsprzedać operatorowi systemu przesyłowego część swoich zobowiązań kontraktowych. I znów, wytwórcy nie muszą konkurować o to, by pasmo ofertowe o charakterze redukcyjnym zostało odsprzedane operatorowi rynku po jeszcze niższej cenie, ponieważ płatność za oferowaną w nim energię bilansującą nastąpi po cenie rozliczeniowej CRO. Jest ona mniejsza od cen ofertowych wszystkich pasm, które nie wejdą do faktycznej produkcji, lub im równa.

Aukcja bilansująca oparta na cenach MCP zapewnia więc generowanie sygnałów ekonomicznych dla uczestników rynku, które skłaniają do kształtowania przedstawianych ofert w sposób odnoszący się bezpośrednio do ponoszonych kosztów. Z punktu widzenia rynku bilansującego, jest to cecha bardzo pożądana. Pamiętajmy bowiem, że rynek ten ma pełnić również (a nawet przede wszystkim) funkcje techniczne, tzn. zapewniać bilansowanie krajowego systemu elektroenergetycznego w optymalny, pod względem kosztów, sposób. Mechanizm MCP zapewnia przywoływanie do produkcji jednostek wytwórczych w kolejności kosztów ich wytwarzania, co pozwala na optymalizację ekonomiczną doboru źródeł energii.

W literaturze (Zerka 2003; Mielczarski 2000) wymienia się jeszcze inne zalety systemu opartego na cenie krańcowej MCP, takie jak: ułatwienie metody rozliczeń, ograniczenie przewagi rynkowej dużych podmiotów biorących udział w rynku w stosunku do mniejszych wytwórców czy też zapewnienie przez rynek wiarygodnej ceny referencyjnej energii elektrycznej.

Należy jednak wspomnieć o pewnej negatywnej właściwości mechanizmu aukcji bilansującej, opartego na cenach MCP. O ile likwiduje on praktycznie grę rynkową uczestników w zakresie określania ceny oferty przy danej cenie

rozliczeniowej CRO, skłaniając ich generalnie do kalkulowania cen ofertowych na poziomie kosztów wytwarzania, o tyle ułatwia tego rodzaju grę przy samym ustalaniu ceny rozliczeniowej CRO. W systemie MCP równowaga rynkowa wyznaczana jest przez jedną, ostatnią ofertę, niezbędną do zrównoważenia popytu. Może to skłaniać wytwórców do stosowania strategii ofertowych w sposób sztuczny podwyższających cenę równowagi, np. przez ograniczanie lub ukrycie części zdolności wytwórczych.

W tego rodzaju strategiach pasma mocy oferty bilansującej, które przekraczają pozycję kontraktową jednostki, organizowane są tak, że jedno, niewielkie, ostatnie pasmo ofertowe otrzymuje bardzo wysoką cenę. Wolumen energii oferowanej w tym paśmie jest na tyle mały, że w przypadku jego odrzucenia i niewejścia do produkcji wytwórca nie ponosi poważnych strat. Jeżeli natomiast zostanie ono przyjęte, winduje cenę całej energii na rynku bilansującym do bardzo wysokiego poziomu. Zagrożenie tego rodzaju jest szczególnie niebezpieczne w przypadku cenotwórców, czyli jednostek wytwórczych o dużej sile rynkowej, które mogą „pomagać" rynkowi w akceptacji takich ofert poprzez ograniczenie swoich zdolności produkcyjnych oferowanych w pasmach o niższych cenach (np. przez fikcyjne awarie, remonty) albo poprzez umiejętne wykorzystanie występujących w pewnych okresach ograniczeń systemowych.

Analizując, z kolei, właściwości aukcji bilansującej opartej na **cenach ofertowych (*pay as bid*)**, należy zauważyć, że w porównaniu z mechanizmem MCP nie zapewnia ona tak dobrych właściwości regulacyjnych, z punktu widzenia optymalizacji wytwarzania w krajowym systemie elektroenergetycznym. Ponieważ rozliczenia następują według cen ofertowych, a nie ceny równowagi, dostawcy usług bilansujących starają się maksymalizować swoje zyski, próbując przewidzieć poziom ceny rynkowej oraz określając swoje ceny ofertowe pasm mocy na podstawie sporządzanych prognoz.

Krzywa podażowa nie odzwierciedla więc ich kosztów wytwarzania energii elektrycznej. Ponadto ewentualna nietrafność ofert wytwórców, spowodowana błędami prognozy ceny równowagi, skutkować może przyjmowaniem przez rynek pasm energetycznych dla źródeł o wyższych kosztach wytwarzania, kosztem tańszych źródeł. W konsekwencji wywołać to może nieefektywność ekonomiczną w dysponowaniu źródłami wytwarzania i wzrost ceny rynkowej.

W literaturze wskazuje się jeszcze inne negatywne właściwości rozliczeń na rynku bilansującym na podstawie mechanizmu cen ofertowych (Zerka 2003; Mielczarski 2000). Wymienić można tu przede wszystkim konieczność wdrożenia efektywnych systemów prognozowania i zarządzania ryzykiem wynikającym z niepewności prognoz ceny równowagi rynku, które stanowią relatywnie większe obciążenie dla małych firm. Ponadto w systemie ofertowym bardziej skomplikowane są rozliczenia i monitorowanie rynku przez organa regulacyjne.

W Polsce w początkowym okresie funkcjonowania rynku bilansującego obowiązywał system rozliczeń oparty na cenach ofertowych. Dosyć powszechnie

uznawano to rozwiązanie za jedną z najważniejszych przyczyn (obok rozwarstwienia cen rozliczeniowych odchyleń, o czym będziemy mówili za chwilę) problemów pojawiających się w działaniu hurtowego rynku energii elektrycznej. W związku z tym od roku 2009 mechanizm rozliczeń na rynku bilansującym został zmieniony na system oparty na cenie krańcowej.

Dokładniej rzecz biorąc, cena rozliczeniowa CRO_h, dla danej godziny handlowej h, równa jest najwyższej cenie za wytwarzanie energii elektrycznej niezbędnej do pokrycia pasmami zdolności wytwórczych aktywnych jednostek grafikowych biorących udział w bilansowaniu zapotrzebowania na energię w obszarze rynku bilansującego (PSE-IRiESP 2012, p. 5.3.1.3.4.3 *Zasady wyznaczania cen rozliczeniowych odchylenia*).

Dla jednostek biorących udział w bilansowaniu rynku rozliczenie różnicy (odchylenia) między ich zadeklarowaną w ZUSE pozycją kontraktową a energią przyjętą przez rynek w pasmach mocy wytwórczych oferty bilansującej (tzw. pozycją kontraktową skorygowaną), czyli tzw. energii bilansującej planowanej, dokonywane jest po cenie rozliczeniowej CRO_h. Jeżeli jednak skorygowana pozycja kontraktowa jest wyższa od pozycji zadeklarowanej, to operator rynku bilansującego płaci wytwórcy za dodatkową energię zakupioną u niego po cenie CRO_h w dodatkowych pasmach (przyrostowych). W sytuacji odwrotnej, gdy pozycja skorygowana jest niższa od zadeklarowanej, okazuje się, że nie wszystkie pasma energii zakontraktowane u wytwórcy przez odbiorców w innych segmentach rynku (kontrakty dwustronne, giełda itp.) zostały przyjęte do fizycznej realizacji. W związku z tym to wytwórca płaci operatorowi rynku bilansującego, kupując brakującą energię po cenie rozliczeniowej CRO_h (oczywiście po to, aby sprzedać ją odbiorcom po określonych wcześniej cenach kontraktowych lub giełdowych).

Podstawową funkcją rynku bilansującego jest wykorzystywanie energii bilansującej (planowanej), zakupionej od dostawców usług bilansujących (aktywnych jednostek grafikowych), do wyrównywania nieplanowanych odchyleń między rzeczywistymi dostawami energii elektrycznej a skorygowaną pozycją kontraktową uczestników hurtowego rynku energii. Jak wspomnieliśmy, w przypadku aktywnych jednostek grafikowych skorygowana pozycja kontraktowa wynika z przyjętych pasm zdolności wytwórczych w ofercie bilansującej. Dla uczestników rynku niebiorących udziału w bilansowaniu, pasywnych (np. odbiorców) pozycja skorygowana odpowiada zadeklarowanej pozycji kontraktowej, zgłoszonej w ZUSE.

Rozliczenia odchyleń nieplanowanych na rynku bilansującym w Polsce również odbywają się na podstawie ceny rozliczeniowej CRO_h, przy czym, oczywiście, niezbilansowanie dodatnie (rzeczywista dostawa wyższa od pozycji kontraktowej) oznacza konieczność zakupu energii elektrycznej od operatora systemu przesyłowego, zaś niezbilansowanie ujemne – konieczność sprzedaży energii.

Regulamin rynku, czyli *Instrukcja ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi* (PSE-IRiESP 2012), daje operatorowi rynku bilansującego możliwość wprowadzenia rozchylenia cen sprzedaży i zakupu energii przez rynek bilansujący poprzez wprowadzenie dwóch cen rozliczeniowych CRO_{Sh}, CRO_{Zh}. Cena sprzedaży przez rynek CRO_{Sh} dla danej godziny handlowej h określana jest jako suma ceny rozliczeniowej odchylenia w tej godzinie (CRO_h) oraz składnika bilansującego ΔB:

$$CRO_{Sh} = CRO_h + \Delta B \tag{1.2.1}$$

zaś cena zakupu CRO_{Zh} określana jest jako różnica tych elementów:

$$CRO_{Zh} = CRO_h - \Delta B \tag{1.2.2}$$

Wartość składnika bilansującego ΔB wyznaczana jest na podstawie różnicy pomiędzy średnią ceną energii na rynku energii elektrycznej, z wyłączeniem centralnego mechanizmu bilansowania handlowego, oraz średnią ceną rozliczeniową odchylenia (CRO). Składnik ten wyznacza się, jeżeli

dla zapewnienia warunków konkurencji na rynku energii elektrycznej lub bezpieczeństwa pracy Krajowego Systemu Elektroenergetycznego jest wymagane tworzenie zachęt ekonomicznych, dla podmiotów uczestniczących w rynku energii elektrycznej, do bilansowania energii elektrycznej dostarczonej i pobranej z systemu w ramach umów sprzedaży energii elektrycznej zawieranych przez te podmioty (PSE-IRiESP 2012, p. 5.3.1.3.4.3.9).

W przeciwnym przypadku składnik ten ma przyjmować wartość 0.

W niezbyt długiej historii polskiego rynku bilansującego obowiązywały już różne uregulowania w tej dziedzinie: począwszy od dosyć poważnego zróżnicowania wartości cen CRO_{Sh}, CRO_{Zh}, poprzez zasadę opartą na korytarzu, tj. niezbilansowania w korytarzu ±1% rozliczane były po cenie rozliczeniowej CRO_h, dopiero różnice wykraczające poza niego – po rozchylonych cenach CRO_{Sh}, CRO_{Zh}. W chwili obecnej, od roku 2009, wskaźnik bilansujący ΔB równy jest zeru, czyli, jak już wcześniej wspomnieliśmy, różnice między pozycją kontraktową a rzeczywistym poborem, rozliczane są po jednolitej cenie CRO_h (tj. $CRO_{Sh} = CRO_{Zh} = CRO_h$).

1.3. Podsumowanie

W bieżącym rozdziale wskazaliśmy źródła i charakter niepewności popytowej na rynku energii elektrycznej. Specyficzne cechy tego towaru, brak możliwości magazynowania energii, konieczność nieustannego równoważenia podaży

i popytu powodują, że niepewność odnośnie do wielkości zapotrzebowania końcowych użytkowników energii przekłada się na generalną niepewność co do wolumenu fizycznej realizacji każdej transakcji na rynku. Biorąc pod uwagę fundamentalny charakter energii elektrycznej dla funkcjonowania społeczeństwa, konieczność zapewnienia bezpieczeństwa zasilania odbiorców, niezbędne jest wbudowanie w strukturę działania rynku energii mechanizmów chroniących przed skutkami tej niepewności.

Funkcję taką pełni rynek bilansujący. Odgrywa on rolę swoistego „magazynu" energii, z którego dostawca pobiera ją w przypadku niemożności zaspokojenia zapotrzebowania odbiorców lub składuje jej nadmiary przy zbyt niskim zapotrzebowaniu. Oczywiście korzystanie z rynku bilansującego niesie za sobą określone ryzyko dodatkowych kosztów związanych z transakcjami bilansującymi. Z tego faktu wynika znaczenie dokładnego i precyzyjnego prognozowania zapotrzebowania na energię. W sytuacji modelowej – gdybyśmy potrafili dokładnie przewidzieć zachowanie odbiorców – moglibyśmy zredukować ryzyko partycypacji w rynku bilansującym do zera. Im dokładniejsza prognoza, tym ryzyko to jest mniejsze. Stanowi to jeden z istotnych czynników leżących u podstaw nieustannej presji na opracowywanie i wprowadzanie do praktycznego zastosowania coraz lepszych metod prognostycznych.

Dążenie do zwiększania dokładności prognozy zapotrzebowania na energię to tylko jeden z aspektów omawianej sytuacji. Pamiętajmy bowiem, że przyszłości nigdy niemal nie da się przewidzieć w sposób idealny; że prognoza nigdy nie jest w pełni dokładna. Wykorzystanie modeli prognostycznych pozwala jedynie na ograniczenie naszej niepewności odnośnie do wielkości zapotrzebowania na energię, nie pozwala natomiast na jej redukcję do zera. Zwróćmy jeszcze raz uwagę, że, jak wskazywaliśmy to w bieżącym rozdziale, niepewność ta dotyczy w zasadzie każdej transakcji na rynku energii. W związku z tym abyśmy mogli szacować ryzyko decyzji związanych z obrotem energią elektryczną, abyśmy mogli włączyć to ryzyko w proces ich podejmowania, oprócz samej prognozy, niezbędne jest modelowanie jej niepewności.

ROZDZIAŁ 2

Metody neuronowe i neuronowo-rozmyte w prognozowaniu krótkoterminowego zapotrzebowania na energię elektryczną

Zgodnie z uwagami poczynionymi w tezie pracy, sieci neuronowe i neuronowo-rozmyte należą obecnie do standardowych narzędzi wykorzystywanych w prognozowaniu krótkoterminowego zapotrzebowania na energię elektryczną. W następnym rozdziale metody te będą przedmiotem szczegółowej analizy, która ma na celu udowodnienie pomocniczej hipotezy badawczej dotyczącej przydatności sposobów oceny wariancji wyjściowej (lub odchylenia standardowego wyjścia) dla tego rodzaju modeli predykcyjnych do szacowania niepewności prognozowanego popytu.

Obecnie natomiast przyjrzymy się zagadnieniom konstrukcji wybranych typów neuronowych i neuronowo-rozmytych narzędzi prognostycznych oraz ich zastosowaniu do predykcji krótkoterminowego zapotrzebowania na energię. Należy jednak pamiętać, że tematem prezentowanej pracy jest modelowanie niepewności prognoz popytu na energię, a nie sam proces ich tworzenia. Szczegółowe zagadnienia związane z technologią prognozowania i wykorzystaniem do tego zadania sieci neuronowych i neuronowo-rozmytych pozostają więc w dużej mierze poza zakresem naszych analiz i zainteresowanych nimi Czytelników odsyłamy do bogatej literatury specjalistycznej.

W rozdziale tym zasygnalizujemy więc tylko najważniejsze zagadnienia z zakresu tworzenia neuronowych i neuronowo-rozmytych modeli prognozy krótkoterminowego zapotrzebowania na energię, koncentrując się dodatkowo na doświadczeniach autora związanych z zastosowaniem tych metod: dokładności uzyskiwanych prognoz i dodatkowych czynników na nie wpływających. Celem tej analizy jest umożliwienie Czytelnikowi lepszego zrozumienia tych metod oceny niepewności działania przedstawianych predyktorów, które zostaną przebadane w następnym rozdziale, by wykazać prawdziwość postawionej w pracy hipotezy badawczej.

Ze względów praktycznych musimy również ograniczyć liczbę omawianych architektur sieci neuronowych i neuronowo-rozmytych wykorzystywanych w zadaniach krótkoterminowego prognozowania zapotrzebowania na energię.

Przedstawiony zostanie jednak dosyć szeroki wachlarz tego rodzaju modeli badanych przez autora, tworzący pewnego rodzaju uporządkowane rodziny architektur. Nie bez znaczenia jest również fakt, że badane architektury sieci mają również największe znaczenie praktyczne dla konstruowania systemów krótkoterminowej prognozy zapotrzebowania na energię i stanowią podstawę wielu zastosowań.

W zakresie sieci neuronowych zajmiemy się przede wszystkim warstwowymi sieciami perceptronowymi (MLP) oraz ich różnego rodzaju wariantami i hybrydami. Można powiedzieć, że charakteryzujemy także i drugą podstawową rodzinę neuronowych modeli prognostycznych, czyli sieci z funkcjami o bazie radialnej (RBF), ponieważ niektóre z przedstawianych dalej sieci neuronowo-rozmytych (sieci FBF) są im funkcjonalnie równoważne. W zakresie modeli neuronowo-rozmytych koncentrujemy się na architekturach tworzących rodzinę tzw. addytywnych systemów z logiką rozmytą, badając po kolei sieci, które realizują wnioskowanie rozmyte dla różnych typów następników reguł: począwszy od stałych następników (sieci FBF), poprzez następniki w formie funkcji liniowych i nieliniowych wejść modelu (wnioskowanie typu Takagi–Sugeno).

2.1. Modelowanie procesu zapotrzebowania na energię

2.1.1. Proces modelowania

Proces modelowania polega na dokonywaniu abstrakcji, czyli świadomego i celowego uproszczonego odwzorowania określonego fragmentu rzeczywistości. Wybór metodologii wykorzystywanej w tym celu związany jest w dużej mierze ze złożonością modelowanego systemu oraz stopniem zrozumienia dynamiki jego zachowania. Ze względu na stopień naszej wiedzy o reprezentowanym fragmencie rzeczywistości możemy mówić o następujących ogólnych rodzajach modeli:

– algorytmiczne,
– dedukcyjne,
– indukcyjne.

Modele algorytmiczne wymagają dogłębnego zrozumienia natury analizowanego zagadnienia i istnienia wiedzy o sposobie jego zachowania, która pozwala na wyspecyfikowanie równań (algorytmu) opisujących dynamikę reprezentowanego systemu w postaci jawnie zdefiniowanych, deterministycznych zależności. Są to tzw. silne modele z silnymi założeniami i bez parametrów wolnych, które należy określić na podstawie zachowania systemu. Stanowią one niewątpliwie najefektywniejszą metodologię rozwiązania problemu, zwykle jednak stosowane mogą być jedynie w przypadku systemów stosunkowo

prostych, dla których możliwe jest precyzyjne zrozumienie i stworzenie opisu modelowanych fenomenów.

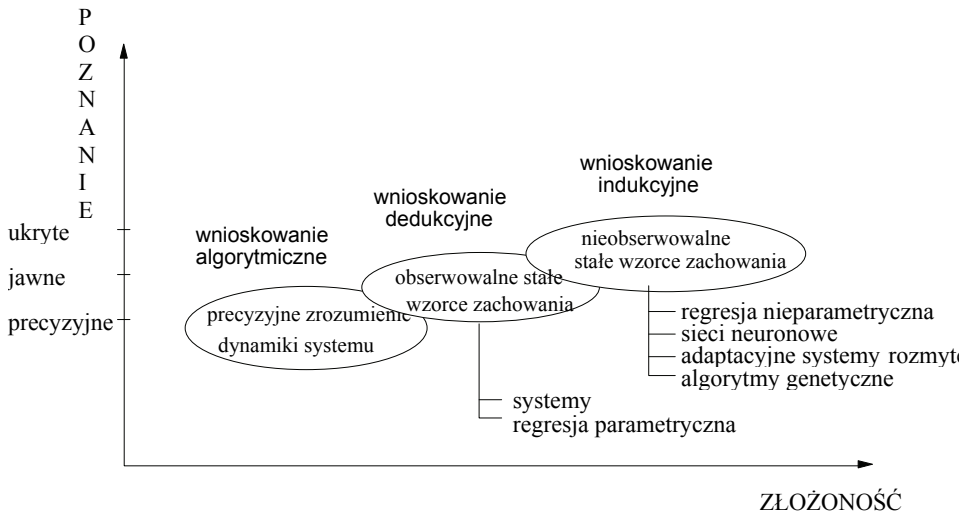
Modele dedukcyjne stosowane są w przypadku systemów, dla których nie jesteśmy w stanie zbudować precyzyjnej specyfikacji matematycznej lub logicznej. Tym niemniej możemy na podstawie bezpośredniej ich obserwacji wykryć pewne stałe wzorce zachowań. Pozwala to na określenie przez modelującego pewnych ogólnych zasad opisujących dynamikę systemu. Typowymi przykładami zastosowania rozumowania dedukcyjnego są systemy ekspertowe czy też metody tradycyjnej analizy statystycznej.

W tym pierwszym przypadku modelujący wraz z ekspertem konstruuje bazę wiedzy opisującą zachowanie systemu w postaci szeregu reguł zachowania, o niewielkiej liczbie (a najczęściej wręcz pozbawionych) parametrów wolnych. Dla odmiany, w systemach wnioskowania statystycznego, takich jak regresja liniowa, czyni się silne założenia odnośnie do natury związku między zmiennymi, pozwalając na oszacowanie parametrów wolnych na podstawie zaobserwowanych danych.

Modele indukcyjne. W miarę wzrostu złożoności systemu możliwość bezpośredniego, precyzyjnego określenia pewnych stałych wzorców jego zachowania zwykle maleje. Powiązania między zjawiskami stają się niejawne, nie można poczynić niemal żadnych założeń odnośnie do ich natury. Problemy tego rodzaju rozwiązywane mogą być przy wykorzystaniu takich metod, jak sieci neuronowe, nieparametryczna regresja, adaptacyjne systemy rozmyte czy też algorytmy genetyczne.

Jak więc widzimy, w miarę wzrostu złożoności modelowanego systemu nasz stopień jego poznania *a priori* zwykle maleje i ciężar przesuwa się w kierunku analizy wzorców jego zachowania na podstawie obserwacji (rys. 2.1.1). Jednocześnie jednak model daje nam coraz mniejsze możliwości poznawcze przy coraz większych nakładach (w szerokim tego słowa znaczeniu) niezbędnych do jego stworzenia.

Jeżeli więc stopień poznania modelowanego systemu pozwala na stosowanie metod „silniejszych", tego typu rozwiązania należy preferować. Powstaje wobec tego pytanie, czy w przypadku modelowania procesu zapotrzebowania na energię niezbędne jest sięganie po złożone metody indukcyjne, takie jak sieci neuronowe czy modele neuronowo-rozmyte. Na to pytanie postaramy się odpowiedzieć w następnym punkcie.



Rysunek 2.1.1. Stopień złożoności systemu a metodologia modelowania
Źródło: opracowanie własne na podstawie A.-P.N. Refenes (ed.), *Neural Networks in the Capital Markets*, Chichester 1995

2.1.2. Charakterystyka procesu zapotrzebowania na energię

Przez **proces zapotrzebowania na energię** rozumiemy proces (sygnał), w którym każdemu wybranemu przedziałowi czasu (zazwyczaj godzina, doba, itp.) przyporządkowana zostaje wartość energii, jaką w nim pobierają odbiorcy. Specyfika systemu elektroenergetycznego spowodowana brakiem możliwości magazynowania energii na skalę przemysłową wymaga, aby energia pobierana przez odbiorców oraz energia sprzedawana przez dostawcę czy też pozyskiwana przez tego dostawcę ze źródeł wewnętrznych i zewnętrznych równoważyły się w każdym momencie. W związku z tym możemy ogólniej określać proces zapotrzebowania na energię jako **proces obciążenia sieci elektroenergetycznej**.

Wśród innych aspektów obciążenia sieci elektroenergetycznej, obok charakterystyki pod kątem energii, istotną informacją zarówno z technologicznego, jak i handlowego punktu widzenia mogą być jego moce chwilowe, a zwłaszcza moc szczytowa (maksymalna) w pewnym zadanym przedziale czasu. Z tego powodu prognozy mocy również mogą należeć do istotnych elementów procesu modelowania zapotrzebowania.

Zapotrzebowanie na energię przez odbiorców w układzie obszarowym (terytorialnym) jest oczywiście procesem (sygnałem) stochastycznym. Do jego modelowania z pewnością nie jesteśmy więc w stanie wykorzystać modeli algorytmicznych. Podstawowe narzędzia stosowane w tej dziedzinie to modele

analizy danych, dedukcyjne (np. regresja liniowa, modele ARIMA) i indukcyjne (np. sieci neuronowe, modele neuronowo-rozmyte).

Podczas analizy procesu krótkoterminowego zapotrzebowania na energię (obciążenia sieci elektroenergetycznej) w układzie terytorialnym, za czynniki wpływające na ten proces przyjmuje się zwykle (Zieliński 2000):

– historyczne wartości obciążeń sieci – mogą być to wartości energii w różnych okresach (godzinowe, dobowe itd.), ich statystyki (np. średnie, maksima, minima) oraz wartości mocy pobieranej przez klientów w okresach szczytów (porannego lub wieczornego),

– zmienne pogodowe – zwykle wykorzystywana jest temperatura, ale również ciśnienie, opady atmosferyczne, siła wiatru itp.,

– inne czynniki – w niektórych przypadkach systemy energetyczne mogą być wrażliwe również na inne czynniki mające wpływ na ich obciążenie, przykładami tego typu oddziaływań mogą być charakter dnia, na który sporządzana jest prognoza (np. dni świąteczne), ważne wydarzenia telewizyjne, zachowania dużych odbiorców.

W prognozach o dłuższym horyzoncie czasowym należy uwzględnić również wpływ innych czynników, dla przykładu (Malko 1995):

– wskaźniki ekonometryczne dla całego kraju – takie jak dochód narodowy brutto, wskaźnik produkcji przemysłowej itp.,

– wskaźniki demograficzne – np. liczba odbiorców energii w obrębie danego zakładu,

– struktura odbiorców – głównie w podziale na odbiorców przemysłowych i gospodarstwa domowe,

– dodatkowe zmienne charakteryzujące odbiorców – przykładowo liczba elektrycznych instalacji kuchennych w danym rejonie, liczba elektrycznych instalacji grzewczych oraz chłodzących itp.

W naszych rozważaniach koncentrujemy się na aspektach krótkoterminowych, więc interesować nas będzie przede wszystkim zależność procesu zapotrzebowania na energię elektryczną od pierwszej grupy czynników, a przede wszystkim od wartości tego procesu z przeszłości oraz od czynników pogodowych, zwłaszcza od temperatur. Szczegółowa analiza procesu krótkookresowego zapotrzebowania na energię przeprowadzona została w publikacji Bartkiewicz 1998b. Tutaj przedstawimy tylko najważniejsze problemy i wnioski dotyczące tego zagadnienia.

Analizując w procesie zapotrzebowania na energię (lub moc) zależności między wartościami tego procesu dla różnych punktów czasu, należy stwierdzić występowanie silnych zależności korelacyjnych (liniowych) między zapotrzebowaniem na energię w okresie bieżącym (prognozowanym) a odpowiadającymi mu wartościami tego procesu z przeszłości. Określając zatem dane wejściowe do modelu prognozy zapotrzebowania na energię, możemy wybrać informacje o opóźnionych obciążeniach sieci za pomocą klasycznej analizy korelacyjnej

i współczynnika determinacji. Dokonując tego wyboru, należy jednak wziąć pod uwagę zmienność sezonową procesu obciążenia sieci. W przypadku prognozy o krótkim horyzoncie czasowym przede wszystkim musimy uwzględnić cykle sezonowe o okresach dobowych i tygodniowych.

Dobowa zmienność zapotrzebowania na energię ma charakterystyczny przebieg, z niższym zapotrzebowaniem w godzinach nocnych oraz wyższym w ciągu dnia, a także z maksymalnymi wartościami w szczycie porannym i wieczornym. Odgrywa ona istotną rolę przy wyborze zmiennych wejściowych prognozy obrotu energią w danej godzinie rozliczeniowej czy też mocy szczytowej. Efekt tej zmienności kompensuje się zazwyczaj, wybierając na wejściu modelu informacje o obciążeniach z godzin zbliżonych do prognozowanej. Pamiętać również należy, że zarówno efekt wartości szczytów porannych i wieczornych, jak i ich umiejscowienie w cyklu dobowym w dłuższym okresie czasu mogą ulegać zmianie. Jest to istotny fakt, który wpływa na konieczność okresowej przebudowy modelu prognostycznego, tak by dopasować go do nowej sytuacji, jak również, ogólnie, na wybór zakresu czasowego wzorców obserwacji historycznych wykorzystywanych do stworzenia modelu.

Problem zmienności tygodniowej zazwyczaj rozwiązywany jest poprzez dodanie sezonowej zmiennej wejściowej (lub zmiennych wejściowych) kodującej dzień tygodnia, na który sporządzana jest prognoza. Dodatkowo czasami podawane są na wejściu informacje o obciążeniach z tego samego dnia, z opóźnieniem tygodniowym.

Podsumowując, informacje o historycznych (opóźnionych) wartościach zapotrzebowania na energię z wybranych okresów stanowią podstawowe dane do sporządzenia krótkoterminowej prognozy obciążeń sieci. Jak już nadmieniališmy, zależności między zmiennymi wejściowymi a wyjściową modelu mają w tym przypadku silny charakter liniowy. W związku z tym jeżeli prognoza opiera się wyłącznie na historycznych wartościach obciążenia sieci, zaleca się stosowanie metod dedukcyjnych, opartych na modelach analizy regresji liniowej.

Tabela 2.1.1. Porównanie współczynników korelacji oraz stosunków korelacyjnych dla danych z poszczególnych kwartałów

Kwartał	$r_{ZD,TMIN}$	$\eta_{ZD,TMIN}$	$r_{ZD,TMAX}$	$\eta_{ZD,TMAX}$
I	–0,33	0,49	–0,14	0,45
II	–0,60	0,74	–0,65	0,83
III	–0,65	0,82	–0,69	0,77
IV	–0,61	0,73	–0,68	0,75

Źródło: opracowanie własne.

Drugą grupą podstawowych czynników wpływających na zapotrzebowanie na energię w krótkim horyzoncie czasowym stanowią warunki pogodowe.

Zazwyczaj wyrażane są one z wykorzystaniem zmiennych określających wartość temperatury w rozważanym okresie prognozy. W przypadku zależności między obciążeniem sieci a temperaturą mamy do czynienia z nieco odmienną sytuacją. W tabeli 2.1.1 zaprezentowane zostały wyniki pomiarów korelacji $r_{ZD,TMIN}$ oraz $r_{ZD,TMAX}$, pomiędzy zapotrzebowaniem dobowym na energię (*ZD*) a temperaturą maksymalną (*TMAX*) oraz minimalną (*TMIN*) w danym dniu. Analizę przeprowadzono dla trzyletniego zbioru danych z systemu jednej ze spółek dystrybucyjnych.

Na podstawie wartości współczynników korelacji podanych w tej tabeli widzimy, że istnieje zależność pomiędzy zapotrzebowaniem na energię a temperaturami. Dokładniejsze badania i testy wskazują, że jest ona istotna statystycznie (patrz Bartkiewicz 1998b).

W tabeli 2.1.1 zaprezentowane zostały również wyniki pomiarów stosunków korelacyjnych $\eta_{ZD,TMIN}$ i $\eta_{ZD,TMAX}$ dla zapotrzebowania na energię i temperatur, wyznaczone na tym samym zbiorze danych. Stosunek korelacyjny zmiennych *Y* oraz *X*, tzn. $\eta_{Y,X}$:

$$\eta_{YX} = \sqrt{\frac{E[m_2(X) - E(Y)]^2}{D^2(Y)}} = \sqrt{1 - \frac{E[Y - m_2(X)]^2}{D^2(Y)}} \tag{2.1.1}$$

przyjmuje wartości z przedziału [0, 1], przy czym 0 oznacza brak zależności między zmiennymi, 1 – istnienie dokładnej zależności funkcyjnej ($m_2(X)$ jest funkcją regresji I rodzaju zmiennej *Y* względem zmiennej *X*). Stosunki korelacyjne stanowią przy tym oszacowanie siły nieliniowej zależności między zmiennymi. Na podstawie tabeli 2.1.1 widzimy więc, że siła zależności nieliniowej w poszczególnych kwartałach jest wyraźnie wyższa niż zależności liniowej. Dokładniejsze analizy przeprowadzone we wcześniejszej publikacji (Bartkiewicz 1998b) również wskazują na istotny statystycznie charakter tej różnicy.

Podsumowując, możemy stwierdzić, że zależność pomiędzy temperaturą a zapotrzebowaniem na energię ma charakter nieliniowy. Zależność ta nie przybiera przy tym żadnego możliwego z góry do wykrycia określonego kształtu, co pozwalałoby na opracowanie transformacji doprowadzającej do modelu krzywoliniowego. Musi więc ona zostać określona na podstawie danych w trakcie oszacowania parametrów modelu. Dlatego w przypadku prognoz zapotrzebowania na energię (obciążenia sieci), w których wykorzystuje się informacje o warunkach pogodowych, niezbędne jest zastosowanie indukcyjnych technik modelowania, takich jak omawiane w naszej pracy sieci neuronowe czy modele neuronowo-rozmyte.

Trzecią wreszcie grupę czynników stanowią nieregularne czynniki mające wpływ na obciążenie sieci, takie jak dni świąteczne, nieregularne cykle dużych

odbiorców itp. Problem uwzględnienia tego rodzaju informacji chwilowo pozostawimy na boku. Ich modelowanie jest zadaniem niełatwym z powodu małej liczby obserwacji i trudności w gromadzeniu danych. Dlatego uwzględniane są one jedynie w niektórych modelach prognozy obciążeń sieci opisywanych w literaturze. Do zagadnienia tego wrócimy jeszcze w punkcie 2.2.3, w którym omawiamy zastosowania sieci neuronowych do konkretnych prognoz.

2.2. Prognozowanie zapotrzebowania na energię z wykorzystaniem modeli neuronowych

2.2.1. Sztuczne sieci neuronowe

Sztuczne sieci neuronowe (SSN) stanowią jedną z najbardziej dynamicznie rozwijających się obecnie technik indukcyjnego modelowania danych. Inspiracją do skonstruowania tej klasy systemów była budowa mózgu ludzkiego. Ten skomplikowany układ, gromadzący i przetwarzający informację, w wielu dziedzinach działa lepiej i sprawniej od najlepszych nawet komputerów. Struktura sieci neuronowej oraz sposób rozwiązywania przez nią zadań przypominają zasadę działania systemu nerwowego. Należy jednak zauważyć, że inspiracje biologiczne, aczkolwiek istotne, dotyczą jedynie ogólnych zasad funkcjonowania SSN. W rzeczywistości działanie większości modeli sieci neuronowych oparte jest na czysto pragmatycznych koncepcjach matematycznych, dostosowanych do rozwiązywanego zadania i mających niewiele wspólnego ze swoimi neurologicznymi podstawami.

Sztuczna sieć neuronowa jest systemem wzajemnie połączonych prostych elementów przetwarzających informacje, zwanych neuronami, jednostkami lub węzłami. Połączenia między elementami mają przyporządkowane współczynniki wagowe, wyznaczające siłę powiązań i tworzące zbiór parametrów modelu. Sieci neuronowej nadaje się zwykle pewną strukturę. Jej jednostki grupowane są w większe zespoły zwane warstwami. Struktura wewnętrzna, wraz z określeniem sposobu propagacji sygnału między neuronami, tworzą tzw. architekturę sieci neuronowej (Jabłoński, Bartkiewicz 2006).

Cała wiedza sieci o sposobie rozwiązania danego problemu przechowywana jest w jej wewnętrznych odwzorowaniach definiowanych przez wartości wag i może być przywołana w procesie reakcji na określony sygnał. Współczynniki wagowe są przydzielone albo wyznaczone w procesie treningowym zmierzającym do nauczenia SSN identyfikowania wzorców albo odwzorowania przekształceń. Zasadniczo więc tworzenie modelu problemu przez sieć neuronową odbywa się na drodze estymacji parametrycznej – oszacowania wag (parametrów) sieci. Układ odwzorowujący SNN jest jednak na tyle bogaty, że sieć może

wykonywać estymację strukturalną i modelować złożone zależności o charakterze nieliniowym bez przyjmowania wcześniejszych założeń o kształcie tych zależności.

Sieci neuronowe realizują najczęściej następujące rodzaje przetwarzania (Zieliński 2000):

– przypominanie polegające na odzyskiwaniu (albo interpretowaniu) zmagazynowanych w SSN informacji, obliczaniu wyjścia dla danego wejścia,

– skojarzenie, które może być realizowane w następujących wariantach: skojarzenie uszkodzonego (zdeformowanego) wejścia (albo wywołania) z najbliższym przechowywanym wzorcem, skojarzenie między parą wzorców, diagnostyka, analiza,

– klasyfikacja, która realizowana jest poprzez podział zbioru wejściowego na klasy lub kategorie i skojarzenie każdego wejścia z kategorią (klasy są zwykle przedstawiane za pomocą dyskretnych wartości wektorów wejściowych, a wyjścia są binarne),

– rozpoznawanie rozumiane jako klasyfikowanie wejścia pomimo tego, że nie odpowiada ono żadnemu z przechowywanych wzorców,

– estymacja, czyli realizacja następujących zadań: aproksymacja, interpolacja, filtrowanie, predykcja, prognozowanie,

– optymalizacja, w tym rozwiązywanie liniowych i nieliniowych równań,

– sterowanie realizowane inteligentnie bez konieczności opracowania modelu, oparte wyłącznie na doświadczeniu.

Model rozwiązania budowany jest przez SNN w procesie uczenia (treningu) sieci, na podstawie dostarczonych tzw. danych treningowych. Polega on na modyfikacji (najczęściej w procesie iteracyjnym) współczynników wagowych połączeń jej elementów. Ze względu na sposób prowadzenia treningu wyróżnić można dwie grupy algorytmów uczących (Hertz, Krogh, Palmer 1993):

1. Uczenie nadzorowane (z nauczycielem) – dane treningowe zawierają zestaw sygnałów wejściowych sieci oraz poprawnych na nie reakcji. Uczenie polega na takiej modyfikacji wag, aby rzeczywiste wyjścia były jak najbliższe wartościom pożądanym. Jeżeli w czasie treningu nie prezentujemy sieci dokładnej wartości pożądanego wyjścia, a jedynie informację, czy reaguje ona prawidłowo, to mamy do czynienia ze specjalnym przypadkiem uczenia nadzorowanego – tzw. uczeniem ze wzmocnieniem.

2. Uczenie bez nadzoru – w procesie uczenia sieć neuronowa nie otrzymuje żadnej informacji na temat pożądanych reakcji. Dane treningowe obejmują jedynie zbiór sygnałów wejściowych. Sieć ma za zadanie samodzielnie zanalizować zależności i korelacje w zbiorze treningowym. Tego typu sieci nazywamy samoorganizującymi (*selforganising networks*) lub autoasocjacyjnymi.

Inną istotną klasyfikację sieci neuronowych przeprowadzić można ze względu na ich architekturę. Wyróżnia się tu trzy główne grupy (Korbicz, Obuchowicz, Uciński 1994):

1. Sieci jednokierunkowe. Ogólnie można powiedzieć, że ich struktura stanowi acykliczny graf skierowany; sieci te mają wyraźnie wyróżnione neurony wejściowe (przyjmujące informacje z zewnątrz) i wyjściowe (przesyłające przetworzoną informację na zewnątrz). Sygnał przekazywany jest zawsze do przodu: z warstwy wejściowej, poprzez jednostki ukryte, do warstwy wyjściowej, bez rekurencyjnych połączeń wstecznych. Dla dowolnego neuronu wartości wejść nie zależą w żaden sposób (bezpośredni czy też pośredni) od jego stanu, czyli wartości wyjściowej. Typowym przykładem takiej sieci jest omawiana dalej dokładnie wielowarstwowa sieć perceptronowa.

2. Sieci rekurencyjne. W przeciwieństwie do sieci jednokierunkowych dopuszczamy występowanie w nich cykli, sygnał wyjściowy neuronu może więc bezpośrednio lub za pośrednictwem innych węzłów być przekazywany na jego wejście. Dynamika działania tego typu sieci jest znacznie bardziej skomplikowana niż w przypadku sieci jednokierunkowych; w sieci rekurencyjnej jednokrotne pobudzenie sieci poprzez sygnał wejściowy powoduje wielokrotną aktywację wszystkich lub tylko części neuronów w procesie tzw. relaksacji sieci. By zapewnić jej poprawne działanie, należy więc spełnić dodatkowy warunek stabilności; pobudzona sieć w skończonym czasie musi osiągać stan stabilny, w którym wartości neuronów dla danego wejścia pozostają stałe; dopiero wówczas określić można wartość wyjścia. Przykładem sieci rekurencyjnej mogą być sieci Hopfielda.

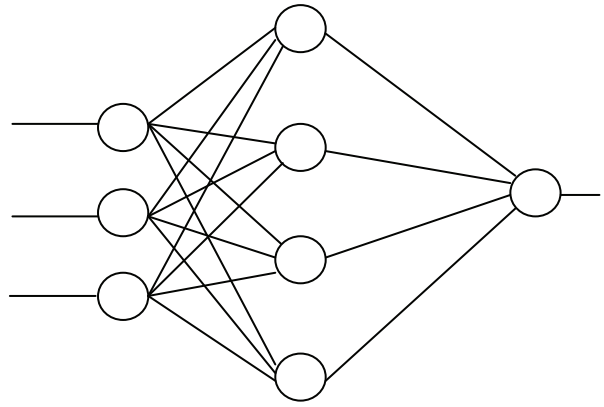
3. Sieci komórkowe. W tej grupie sieci neuronowych wprowadza się dodatkowo pojęcie sąsiedztwa węzłów. Połączone między sobą są tylko jednostki znajdujące się w jego obrębie; charakter tych powiązań może być różny, zależny od konkretnego przypadku. Przykładem tego typu sieci mogą być neuronowe sieci komórkowe (*cellural neural networks*); do tej kategorii zaliczyć można również sieci SOM Kohonena.

2.2.2. Warstwowe sieci perceptronowe

Warstwowe sieci perceptronowe (*Multilayered Prerceptrons*, MLP) są przykładem jednokierunkowych sieci, zazwyczaj o nieliniowej charakterystyce przetwarzania. Stanowią one niewątpliwie jedną z najczęściej wykorzystywanych w praktycznych aplikacjach architektur SNN, zwłaszcza w zastosowaniach związanych z identyfikacją systemów, estymacją i klasyfikacją. Stosunkowo prosta i dobrze poznana specyfika działania tej sieci oraz jej bardzo interesujące własności aproksymacyjne powodują, że projektanci systemów chętnie sięgają po to narzędzie, zarówno w modelach badawczych, jak i w aplikacjach komercyjnych.

Architektura, sposób działania i algorytmy uczenia warstwowych sieci perceptronowych były już wielokrotnie opisywane w szeroko dostępnej literaturze przedmiotu (odsyłamy tutaj czytelnika do takich pozycji jak np. Hertz, Krogh, Palmer 1993; Korbicz, Obuchowicz, Uciński 1994; Masters 1996; Żurada,

Barski, Jędruch 1996; Zieliński 2000), w związku z tym obecnie przedstawimy jedynie najważniejsze zagadnienia związane z tym tematem, przechodząc do kwestii zastosowań sieci perceptronowych w krótkoterminowych prognozach obciążeń sieci elektroenergetycznej.



Rysunek 2.2.1. Warstwowa sieć perceptronowa (MLP) o strukturze {3, 4, 1}, trzech neuronach wejściowych, jednym neuronie wyjściowym i czterech w pojedynczej warstwie ukrytej
Źródło: opracowanie własne

Jak już wspomnieliśmy, MLP stanowią przykład sieci jednokierunkowych. Neurony zgrupowane są w co najmniej dwu warstwach. Pierwsza z nich nazywana jest warstwą wejściową, ostatnia zaś warstwą wyjściową. Między nimi wystąpić może jedna lub więcej warstw ukrytych (rys. 2.2.1). Sygnał przekazywany może być jedynie „do przodu", a więc z warstwy poprzedniej do następnej. Niemożliwe jest przekazywanie sygnałów między neuronami tej samej warstwy lub wstecz, np. z warstwy wyjściowej do ukrytej.

W omawianych w bieżącym rozdziale modelach wykorzystywano sieci MLP z jedną warstwą ukrytą i jednym neuronem w warstwie wyjściowej. Równanie przetwarzające takiej sieci, dla danego wzorca wejściowego (x_1, x_2, …, x_n), możemy zapisać:

$$y(x_1, x_2, ..., x_n) = \varphi\left(\sum_{i=1}^{h} w_i^{(2)} \varphi\left(\sum_{j=1}^{n} w_{ij}^{(1)} x_j\right)\right) \tag{2.2.1}$$

gdzie h jest liczbą neuronów w warstwie ukrytej, $w_{ij}^{(1)}$ współczynnikiem wagowym j-tego wejścia, i-tego neuronu w warstwie ukrytej, zaś $w_i^{(2)}$ jest współczynnikiem wagowym i-tego wejścia neuronu wyjściowego.

Jako funkcję aktywacji neuronów φ w wykorzystywanych sieciach przyjęto funkcję logistyczną:

$$\varphi(net) = \frac{1}{1 + \exp(net - \Theta))} \tag{2.2.2}$$

gdzie próg Θ zrealizowano jako wagę dodatkowego neuronu roboczego o stałym wyjściu (tzw. bias), zaś *net* oznacza łączne pobudzenie neuronu, czyli iloczyn skalarny wektora wag i wejść neuronu:

$$net = \mathbf{w}^T \mathbf{x} = \sum_i x_i w_i \tag{2.2.3}$$

Do uczenia sieci zastosowano standardowy algorytm uczenia sieci MLP, tzw. algorytm wstecznej propagacji błędu. Bardziej szczegółowe przedstawienie wykorzystywanych wzorów znajduje się w załączniku Z1.1. Tutaj przypomnijmy tylko krótko zasadę jego działania. I tak algorytm ten polega na iteracyjnej prezentacji sieci pewnego zbioru treningowego:

$$\{\mathbf{x}_k, t_k\} = \{(x_{k1}, \ldots, x_{kn}), t_k\}, k = 1, \ldots, N \tag{2.2.4}$$

gdzie $\mathbf{x}_k = (x_{k1}, \ldots, x_{kn})$, $k = 1, \ldots, N$ jest pewnym znanym wzorcem wartości wejść, zaś t_k odpowiadającą mu znaną wartością wyjściową (tzw. wartością treningową).

Cel uczenia polega, rzecz jasna, na tym, aby reakcja sieci na każdy z wzorców wejściowych $\mathbf{x}_k$, tj. wyjście sieci $y(x_{k1}, \ldots, x_{kn})$, była jak najbliższa wartości treningowej t_k, w sensie minimalizacji błędu kwadratowego modelu (Zieliński 2000):

$$E = \frac{1}{2}\sum_k (t_k - y(x_{k1},..,x_{kn}))^2 \tag{2.2.5}$$

W algorytmie wstecznej propagacji błędu, w jego klasycznej postaci, wykorzystuje się do minimalizacji błędu metodę najszybszego spadku. Modyfikacje parametrów modelu, czyli wag, powinny być dokonywane w kierunku przeciwnym do wektora gradientu funkcji błędu, a zatem:

$$\Delta w = -\eta \frac{\partial E}{\partial w} \tag{2.2.6}$$

dla każdej wagi sieci *w*. Współczynnik η reguluje długość kroku modyfikacji i zwany jest współczynnikiem uczenia.

Po zróżniczkowaniu funkcji błędu względem każdej z wag sieci algorytm wstecznej propagacji błędu daje następującą regułę na modyfikację wartości tych wag (patrz załącznik Z1.1, zależność (Z1.8)):

$$\begin{aligned} \Delta w_i^{(2)} &= \eta \delta^{(2)} y_i \\ \Delta w_{ij}^{(1)} &= \eta \delta_i^{(1)} x_{kj} \end{aligned} \tag{2.2.7}$$

gdzie oznaczenia są takie same jak we wzorze (2.2.1) oraz y_i jest stanem (wyjściem) i-tego neuronu warstwy ukrytej (czyli i-tym wejściem neuronu wyjściowego), zaś δ błędami poszczególnych neuronów, obliczanymi według wzoru (patrz (Z1.10) i (Z1.12)):

$$\begin{aligned} \delta^{(2)} &= \varphi'(net^{(2)})(t_k - y(x_1,...,x_n)) \\ \delta_i^{(1)} &= \varphi'(net_i^{(1)}) w_i^{(2)} \delta^{(2)} \end{aligned} \tag{2.2.8}$$

Zauważmy jeszcze tylko, że pochodną logistycznej funkcji aktywacji możemy łatwo wyznaczyć ze wzoru:

$$\varphi'(net) = \varphi(net)(1 - \varphi(net)) \tag{2.2.9}$$

2.2.3. Prognozowanie dobowego zapotrzebowania na energię z wyprzedzeniem jednodniowym przy wykorzystaniu sieci MLP

Jako pierwszy przypadek zastosowania modeli neuronowych przeanalizujemy zagadnienie predykcji dobowego zapotrzebowania na energię z jednodniowym wyprzedzeniem czasowym. Zgodnie z dyskusją w punkcie 2.1.2, jako zmienne wejściowe prognozy wykorzystane zostały informacje o obciążeniu sieci w dniu poprzednim, temperatury w dniu poprzednim oraz w dniu prognozowanym. Dla uwzględnienia zmienności tygodniowej wprowadzono dodatkowe zmienne kodujące dzień tygodnia. Dokładniej funkcja realizowana przez model ma następującą postać:

$$\begin{aligned} ZD_d = f(&ZG_{d-1}(1),...,ZG_{d-1}(24), TMIN_{d-1}, \\ &TMAX_{d-1}, TMIN_d, TMAX_d, dt_{1d},...,dt_{6d}) \end{aligned} \tag{2.2.10}$$

gdzie:

ZD_d – dobowe zapotrzebowanie na energię w dniu d,
$ZG_{d-1}(1)$, ..., $ZG_{d-1}(24)$ – godzinowy rozkład zużycia w dniu poprzednim,
$TMIN_{d-1}$, $TMAX_{d-1}$ – temperatura minimalna i maksymalna w dniu poprzednim,
$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,
dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

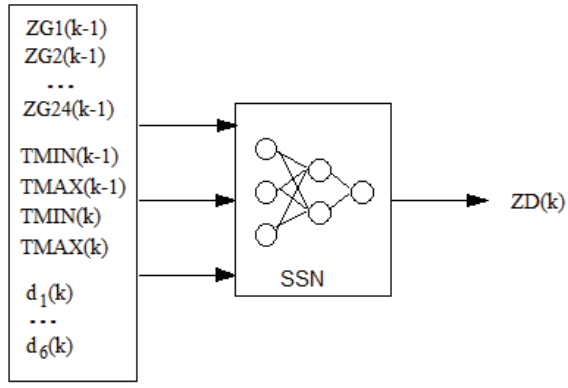
Do zakodowania dnia tygodnia przyjęto standardowy sposób kompensacji sezonowej dla modeli regresji, przy użyciu binarnych zmiennych roboczych:

$$dt_{it} = \begin{cases} 1 & \text{dla obserwacji z } i \text{ - tego dnia tygodnia} \\ 0 & \text{dla obserwacji z innych dni tygodnia} \end{cases} \tag{2.2.11}$$

I tak $dt_{1d} = 1$, jeżeli dzień d to poniedziałek, dla wtorku $dt_{2d} = 1$, dla środy $dt_{3d} = 1$ itd. Niedziela kodowana jest przez zerowe wartości wszystkich zmiennych.

Do modelowania funkcji f wykorzystano warstwową sieć jednokierunkową. Jeden z istotnych czynników w budowie modelu neuronowego stanowi dobór odpowiedniej struktury sieci. Rozmiar warstwy wejściowej i wyjściowej determinowany jest przez liczbę zmiennych zależnych i niezależnych modelu. Koniecznie należy natomiast ustalić liczbę warstw ukrytych oraz liczby neuronów w każdej z nich. Można pokazać (patrz np. Hornik, Stinchcombe, White 1989; Bartkiewicz 1998b), że do aproksymacji dowolnej funkcji ciągłej za pomocą sieci MLP wystarczy jedna warstwa ukryta. Przyjmuje się więc, że sieci MLP posiadają tzw. właściwość uniwersalnej aproksymacji, dlatego należą one do standardowych modeli stosowanych w zagadnieniach prognostycznych. Jeśli występują kłopoty z nauczeniem sieci właściwego odwzorowania, możemy próbować zwiększać siłę aproksymacyjną modelu poprzez dodanie nowych warstw ukrytych.

Pozostaje problem liczby neuronów w warstwie ukrytej. Właściwość uniwersalnej aproksymacji nie daje żadnych wskazówek w tej materii. Zapewnia jedynie, że liczba ta jest skończona. W większości przypadków dobierana jest ona eksperymentalnie. Zaczynamy od sieci o prostej strukturze, a następnie, w przypadku niemożności osiągnięcia dostatecznie dobrej aproksymacji, zwiększamy liczbę neuronów ukrytych.



Rysunek 2.2.2. Struktura neuronowego modelu prognozy dobowego zapotrzebowania na energię
Źródło: opracowanie własne

Istotne przy tym jest, aby architektura sieci nie stała się zbyt bogata. W przypadku dużej liczby neuronów ukrytych sieć za bardzo dopasowuje się do danych, co powoduje pogorszenie jej działania w populacji generalnej (patrz dyskusja o wymienności obciążenia i wariancji w punkcie 3.1.3).

W naszym przypadku do prognozy zastosowano sieć z jedną warstwą ukrytą o strukturze {34, 6, 1} (Bartkiewicz 1998b; Bardzki, Bartkiewicz, Gontar, Zieliński 1998; Zieliński 2000; Bartkiewicz, Czajkowska i inni 2004), gdzie powyższe wartości odpowiadają liczbom neuronów w poszczególnych jej warstwach. Liczba 34 neuronów wejściowych odpowiada liczbie zmiennych objaśniających w modelu (2.2.10), podobnie 1 neuron wyjściowy reprezentuje prognozę pojedynczej zmiennej *ZD* (rysunek 2.2.2). Liczba 6 neuronów w warstwie ukrytej dobrana została na drodze eksperymentalnej, tak by uzyskać model o jak najlepszej generalizacji.

Do uczenia sieci wykorzystano dwuletni zbiór danych z jednej ze spółek dystrybucyjnych. Sieć MLP uczono przy użyciu klasycznego algorytmu wstecznej propagacji błędu, opisanego w punkcie 2.2.2, z dodatkowym członem regularyzacyjnym podczas modyfikacji wag, tzw. momentum. W każdym kroku uczenia n, po obliczeniu za pomocą (2.2.7) poprawki dla danej wagi $\Delta w(n)$, była ona dodatkowo korygowana wartością poprawki z poprzedniego kroku

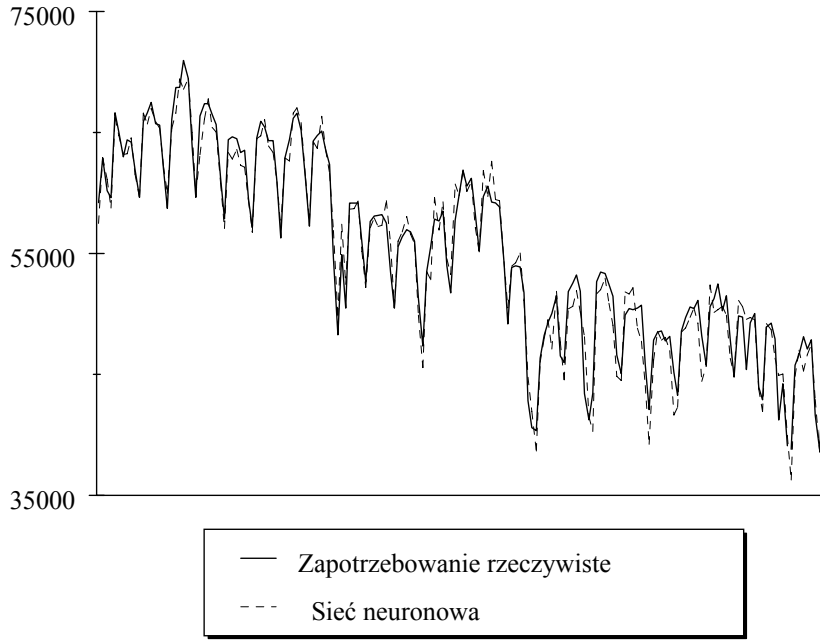
$$\Delta w(n) = \Delta w(n) + \alpha \Delta w(n-1) \tag{2.2.12}$$

gdzie α jest parametrem dobieranym do konkretnego przypadku, w powyższym modelu, na drodze eksperymentalnej, przyjęto $\alpha = 0{,}2$. Regularyzacja przy wykorzystaniu współczynnika momentum jest znaną metodą poprawiania zbieżności algorytmu uczącego pozwalającą na redukcję możliwości przeskoków pomiędzy zboczami powierzchni błędu.

Poprawka dla wag (2.2.7) w metodzie wstecznej propagacji błędu zależna jest od parametru uczenia η. Współczynnik ten również dobierany jest do konkretnego problemu, często na drodze eksperymentalnej. Zaleca się jednak, aby malał on w miarę postępów uczenia sieci. W przypadku analizowanego obecnie modelu zastosowano metodę adaptacyjnego doboru współczynnika w zależności od kształtu powierzchni błędu. Współczynnik η zmienia się w każdym kroku algorytmu uczenia oraz dobierany jest indywidualnie dla każdej wagi w, zgodnie z zależnością:

$$\eta_w(n) = \begin{cases} u\eta_w(n-1) & \text{gdy} \dfrac{\partial E}{\partial w}(n) \cdot \dfrac{\partial E}{\partial w}(n-1) > 0 \\ d\eta_w(n-1) & \text{gdy} \dfrac{\partial E}{\partial w}(n) \cdot \dfrac{\partial E}{\partial w}(n-1) < 0 \end{cases} \tag{2.2.13}$$

gdzie u i d są stałymi dodatnimi, odpowiednio nieco większą i nieco mniejszą od 1, n jest indeksem kroku w metodzie wstecznej propagacji błędu.



Rysunek 2.2.3. Porównanie prognozy i rzeczywistego zapotrzebowania na energię
Źródło: opracowanie własne

Tabela 2.2.1. Porównanie dokładności działania modelu neuronowego i regresji liniowej w okresie testowym

Modele	MAE (MWh)	MAX AE (MWh)	RMSE (MWh)	MAPE (%)	MAX APE (%)
Regresja liniowa	119,4	952,7	176	2,31	22,25
Sieć neuronowa	106,1	470	141	2,04	10,83

Źródło: opracowanie własne.

Model przetestowany został na półrocznym zbiorze danych. Wykres rzeczywistej i prognozowanej wielkości zapotrzebowania na energię w okresie testowym zaprezentowano na rysunku 2.2.3. Jak widzimy, różnice pomiędzy obiema wielkościami na wykresie są minimalne i prognoza dobrze oddaje przebieg procesu zapotrzebowania.

W tabeli 2.2.1 przedstawione zostały wyniki testowania modelu. Jako miary oceny działania predyktora wykorzystano standardowe miary błędów statystycznych:

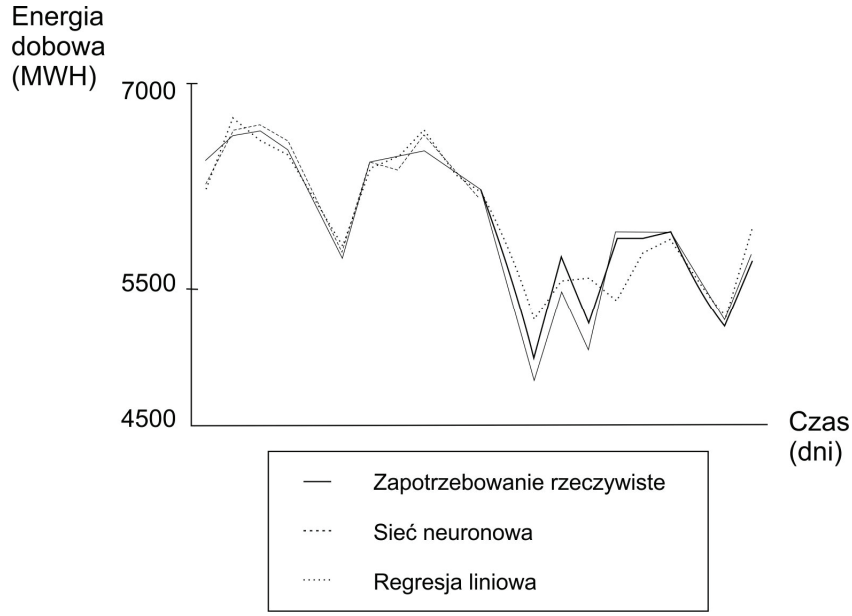
– MAE – średni błąd bezwzględny (*Mean Absolute Error*) – średnia z wartości bezwzględnej odchyleń poszczególnych prognoz,

– MAX AE – maksymalny błąd bezwzględny (*Maximum of Absolute Error*) – największa wartość bezwzględna odchylenia poszczególnych prognoz,

– MAPE – średni bezwzględny błąd procentowy (*Mean Absolute Percentage Error*) – średnia z procentowych wartości bezwzględnych odchyleń poszczególnych prognoz w stosunku do wartości faktycznej zapotrzebowania,

– MAX APE – maksimum bezwzględnego błędu procentowego (*Maximum of Absolute Percentage Error*) – maksymalna wartość bezwzględna odchylenia procentowego.

Analizując wyniki testowania, widzimy, że średni błąd prognozy dla modelu neuronowego wyniósł około 2%, zaś maksymalny jest rzędu 11%. Dla porównania w tabeli 2.2.1 przedstawiono również wyniki działania modelu liniowego. Jak widać, są one wyraźnie gorsze niż w przypadku nieliniowej sieci MLP. Na uwagę zasługują ponadto wartości błędu maksymalnego. Sieć neuronowa jedynie raz w okresie testowym popełniła błąd wyższy niż 10%, przekraczając ten próg tylko nieznacznie. Błąd maksymalny modelu regresji był ponad dwukrotnie większy.



Rysunek 2.2.4. Porównanie działania modelu liniowego i neuronowego dla wybranych dni
Źródło: opracowanie własne

Na rysunku 2.2.4 zobrazowano porównanie działania modelu neuronowego i liniowego dla fragmentu okresu testowego. Jego analiza pozwala rzucić pewne światło na przyczyny lepszego działania sieci neuronowej. Lewa część wykresu

obejmuje okres spokojnych, stałych zmian prognozowanego zapotrzebowania na energię, związanych niemal wyłącznie z tygodniowym wahaniem sezonowym. Prognozowane są one na podstawie historycznych wzorców zachowań, w których występują silne zależności liniowe. Prawa część wykresu natomiast obejmuje okres o zmiennych warunkach pogodowych. Powoduje to powstanie licznych punktów zmiany trendu oraz gwałtownych skoków prognozowanego procesu. Zależności te są wyraźnie lepiej modelowane przez predyktor neuronowy, co potwierdza hipotezę dotyczącą ich nieliniowego charakteru.

2.2.4. Prognozowanie godzinnego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym

Bardziej nawet użyteczną prognozą popytu na energię może być prognoza godzinnego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym. Wiąże się to z omawianą wcześniej organizacją rynku energii elektrycznej w Polsce, a przede wszystkim z działaniem rynku transakcji natychmiastowych dnia następnego, w postaci giełdy energii oraz rynku bilansującego. Cena energii na rynku zmienia się w cyklu godzinnym, z wyprzedzeniem jednodniowym. Niezbędne jest przy tym sporządzenie wcześniejszych grafików zakupów, związane z koniecznością bilansowania niezrównoważenia podaży i popytu na całym rynku poprzez udział w rynku bilansującym.

Planowanie wielkości zakupów wymaga więc od spółki dystrybucyjnej również określania w podobnym cyklu przewidywanego zapotrzebowania jej odbiorców. Szybkie ruchy cen na rynku wymuszają odpowiednią reakcję spółki w dostosowaniu struktury zakupów do zmieniającej się struktury kosztów. Na rynkach dnia następnego proces ofertowy transakcji energią na daną dobę handlową d toczy się zazwyczaj w dniu poprzednim, $d-1$, przy czym różne rynki zamykają się o różnych godzinach, zaś zgłaszanie ZUSE odbywa się do godziny 14.30. Planowanie dostaw spółki obracającej energią musi więc odbywać się na podstawie informacji wcześniejszych, pochodzących z dnia $d-2$. Model prognostyczny wspomagający ten proces wymaga wobec tego dłuższego horyzontu czasowego predykcji i powinien dokonywać predykcji obciążeń co najmniej z wyprzedzeniem dwudniowym.

Na potrzeby wspomagania krótkoterminowego planowania wielkości dostaw w regionalnym systemie elektroenergetycznym spółki dystrybucyjnej (Bardzki, Bartkiewicz, Gontar, Zieliński 1999; Bartkiewicz, Gontar, Zieliński, Bardzki 2000a; Bartkiewicz, Gontar, Matusiak, Zieliński 2001b) stworzono oprogramowanie prognostyczne, w którym zastosowano następujący model predykcji zapotrzebowania na energię:

$$ZG_d(t) = f(ZG(t)_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZG(t)_{d-2}, ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \quad (2.2.14)$$

gdzie oznaczenia są analogiczne jak w (2.2.10):

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Jako dane wejściowe do prognozy zapotrzebowania na daną godzinę wykorzystano więc informacje o obciążeniach sprzed dwóch dni, z tej samej godziny i godzin z nią sąsiadujących oraz z tej samej godziny sprzed tygodnia, prognozy temperatur na dany dzień i zmienne kodujące dzień tygodnia.

Tabela 2.2.2. Błędy prognozy godzinnego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym dla sieci MLP

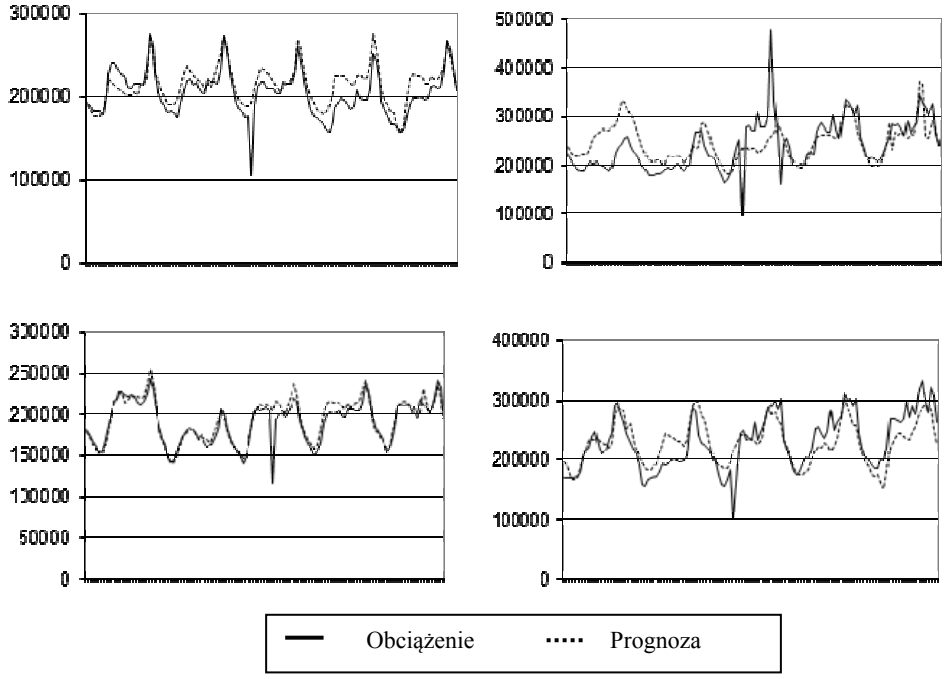
Godz.	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)	Godz.	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
1	7 310	57 157	3,67	35,45	13	11 131	100 016	4,98	92,18
2	6 336	43 998	3,37	24,68	14	12 121	99 222	5,20	43,40
3	5 868	40 208	3,19	22,80	15	11 361	88 634	5,00	76,25
4	5 884	36 017	3,22	20,86	16	13 118	77 893	5,63	40,41
5	6 388	44 036	3,52	24,67	17	13 652	103 047	5,60	48,67
6	7 526	85 901	4,11	81,30	18	13 359	219 901	5,33	45,84
7	9 528	109 284	4,83	107,4	19	13 135	87 038	5,28	33,59
8	10 970	71 606	4,97	34,93	20	12 155	80 425	4,99	43,27
9	9 847	139 669	4,46	146,63	21	11 197	111 810	4,43	69,85
10	10 668	77 782	4,56	39,99	22	10 620	82 074	4,11	33,69
11	10 206	89 878	4,43	46,73	23	8 018	63 362	3,41	28,13
12	10 035	84 254	4,44	47,09	24	7 333	68 560	3,40	37,57
Średnia dla wszystkich godzin						9 907		4,42	
Prognozy dobowego zapotrzebowania na energię						171 000	1 488 000	3,18	29,15

Źródło: opracowanie własne.

Model prognostyczny składał się z 24 równań postaci (2.2.14), z których każde przypisane jest określonej godzinie; odwzorowano go przy zastosowaniu oddzielnych sztucznych sieci neuronowych MLP o strukturze {13, 10, 1} (13 neuronów w warstwie wejściowej, 10 w warstwie ukrytej, 1 neuron wyjściowy) (Bartkiewicz, Gontar, Zieliński, Bardzki 2000a). Rozwiązanie to dało lepsze wyniki niż zastosowanie jednej sieci neuronowej z 24 wyjściami, dającej prognozę całego wzorca przebiegu procesu zapotrzebowania na energię w czasie doby.

W procesie modelowania wykorzystano obserwacje obciążeń sieci i temperatur pochodzące z jednej ze spółek dystrybucyjnych. Dane treningowe obejmowały dwuletni zbiór obserwacji. System prognostyczny przetestowany został na długim, ponadtrzyletnim zbiorze danych. Wykorzystane dane obejmowały wszystkie dni tygodnia, włączając soboty, niedziele i dni świąteczne. Wyniki testowania systemu przedstawione zostały w tabeli 2.2.2.

Jak widzimy w tabeli 2.2.2, otrzymane wyniki prognozy są wyraźnie słabsze niż prezentowane dla modelu (2.2.10) w punkcie 2.2.3. Średni błąd prognozy ukształtował się na poziomie 4,42% (9907 kWh), średnie błędy dla poszczególnych godzin przekraczają nawet w niektórych przypadkach 5%. Charakterystyczne są również duże błędy maksymalne, przekraczające w pewnych okresach doby poziom 100%. Oczywiście należy pamiętać, że proces godzinnego zapotrzebowania na energię charakteryzuje się dużo większą zmiennością względną niż proces zapotrzebowania dobowego, nawet jednak prognozy obciążeń dobowych w tabeli 2.2.2, zsumowane z prognoz godzinnych, cechują się stosunkowo wysokim błędem, który nie do końca można wyjaśnić wydłużeniem horyzontu czasowego prognozy z jednego dnia do dwóch.



Rysunek 2.2.5. Przykłady nagłych fluktuacji procesu obciążenia sieci (zapotrzebowania na energię) w spółce dystrybucyjnej

Źródło: W. Bardzki, W. Bartkiewicz, Z. Gontar, J.S. Zieliński, A *survey of short-term load forecasting algorithms for transient period*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk (eds), *Present-Day Problems of Power Engineering (APE '99)*, vol. 5, Gdańsk–Jurata 1999, s. 11–18

Niektóre z powstałych dużych błędów systemu prognostycznego związane są z nieoczekiwanymi fluktuacjami obciążenia sieci i w związku z tym mają charakter nieunikniony, jako wynik nieodłącznej losowości procesu godzinnego zapotrzebowania na energię w spółce dystrybucyjnej (rysunek 2.2.5). Tego typu przypadki zdarzają się jednak stosunkowo rzadko; związane są one z rozległymi awariami, nieregularnymi procesami produkcyjnymi dużych odbiorców itp. Tym niemniej okazuje się, że duża część błędów związana jest z pewnymi zjawiskami przewidywalnymi, mianowicie nietypowymi wzorcami zachowania procesu zapotrzebowania na energię dla określonych typów dni. Zagadnieniu temu przyjrzymy się dokładniej w następnym punkcie.

2.2.5. Modelowanie dni nietypowych z wykorzystaniem podejścia neuronowo-heurystycznego

Szczegółowa analiza dokładności działania modelu prognostycznego przedstawionego w poprzednim punkcie wskazała, że większość dużych błędów wiąże się z nietypowymi wzorcami zapotrzebowania na energię, które pojawiają się w przypadku dni specjalnych. Pod tym pojęciem rozumiemy przede wszystkim święta narodowe i religijne przypadające w normalne robocze dni tygodnia i w związku z tym nieuwzględniane przez mechanizm modelowania sezonowości tygodniowej, takie jak Wielkanoc (poniedziałek), 3 maja, 1 i 11 listopada, Boże Narodzenie itp. Jedynie kilka spośród najmniej dokładnych prognoz nie dało się wyjaśnić wpływem nietypowych wzorców zachowania odbiorców związanych z dniami specjalnymi.

Zauważmy ponadto, że jako dni specjalne należy uwzględnić także dni poświąteczne i dwa dni po święcie. W prognozach sporządzanych dla tych dni jako dane wejściowe wykorzystuje się zaburzony profil zapotrzebowania na energię z dnia świątecznego. Po odrzuceniu obserwacji dla dni specjalnych ze zbioru testowego średni błąd modelu prognostycznego MLP (2.2.14) dla wszystkich godzin spadł z poziomu 4,42% (tabela 2.2.2), do poziomu 3,76% (tabela 2.2.3).

Podsumowując, dla celów modelowania nietypowych zachowań odbiorców energii w dniach specjalnych wyróżnione zostały więc cztery odmienne typy wzorców zapotrzebowania na energię (Bartkiewicz, Gontar, Matusiak, Zieliński 2002):

– wzorce dla dnia normalnego,
– wzorce dla dnia świątecznego,
– wzorce dla dnia bezpośrednio po dniu świątecznym,
– wzorce dla dnia dwa dni po dniu świątecznym.

Tabela 2.2.3. Średni błąd prognozy MLP dla dni normalnych

MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
8 754	9 998	3,76	43,40

Źródło: opracowanie własne.

Tabela 2.2.4. Średni błąd prognozy dla MLP dni normalnych po usunięciu ze zbiorów treningowych obserwacji dla dni specjalnych

MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
6 951	62 571	3,04	31,40

Źródło: opracowanie własne.

Dla dni normalnych zastosowano predyktor opisany w poprzednim punkcie. Model składał się więc z 24 równań postaci (2.2.14) realizowanych przez warstwowe sieci perceptronowe MLP, o strukturze {13, 10, 1}. W odróżnieniu od poprzedniego przypadku zbiory treningowe dla sieci obejmowały wyłącznie obserwacje dla dni normalnych (Bartkiewicz, Gontar, Zieliński, Bardzki 2000a). Pozwoliło to uniknąć wprowadzenia w procesie uczenia sieci dodatkowych szumów i niejednoznaczności generowanych przez odstające, nietypowe wzorce zachowań odbiorców w dniach specjalnych. Dzięki temu osiągnięto znaczną poprawę dokładności działania modelu dla dni normalnych, redukując jego średni błąd z poziomu 3,76% (tabela 2.2.3) do poziomu 3,04% (tabela 2.2.4).

Model dla dni specjalnych opracowany został przez doktora Zbigniewa Gontara z Katedry Informatyki Uniwersytetu Łódzkiego (Bartkiewicz, Gontar, Zieliński, Bardzki 2000a; Bartkiewicz, Gontar, Zieliński, Bardzki 2000b; Bartkiewicz, Gontar, Matusiak, Zieliński 2002). Podobnie jak w przypadku dni normalnych, zastosowano odrębne równanie dla każdej prognozowanej godziny, ale o nieco zmodyfikowanej postaci:

$$ZG_d(t) = f(ZG(t)_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZG(t)_{d-2}, ZG(t+1)_{d-2}, ZG(sr)_{d-2} TMIN_d, TMAX_d) \tag{2.2.15}$$

gdzie wszystkie oznaczenia są identyczne jak w (2.2.14) oraz $ZG(sr)_d$ jest średnim zapotrzebowaniem na energię w dniu d.

Do oszacowania modelu (2.2.15) wykorzystano warstwowe sieci perceptronowe MLP o strukturze {8, 6, 1}. Liczba obserwacji dla dni specjalnych jest jednak zbyt mała, by wykorzystać je do treningu sieci neuronowej, wynosi

typowo około 25 przypadków rocznie. Z tego powodu święta oraz inne dni specjalne zasymulowane zostały przy użyciu wzorców zapotrzebowania dla niedziel i generalnie dni weekendowych. W tym celu zastosowane zostały pewne dodatkowe wyrównania heurystyczne związane z interpretacją zmiennych wejściowych:

– dla dnia świątecznego do treningu sieci wykorzystane zostały obserwacje dotyczące dni świątecznych i niedziel; $ZG(t)_{d-7}$ oznacza w tym przypadku ostatnią niedzielę, a $ZG(t)_{d-2}$ przedostatni dzień roboczy przed dniem świątecznym,

– dla dnia poświątecznego do treningu sieci wykorzystane zostały obserwacje dotyczące dni poświątecznych i poniedziałków; $ZG(t)_{d-7}$ oznacza w tym przypadku ostatni poniedziałek, a $ZG(t)_{d-2}$ – ostatni dzień roboczy przed dniem świątecznym,

– dla dnia następującego dwa dni po dniu świątecznym wykorzystane zostały obserwacje dotyczące takich dni i wtorków; $ZG(t)_{d-7}$ oznacza w tym przypadku ostatni wtorek, a $ZG(t)_{d-2}$ – ostatni dzień roboczy przed dniem świątecznym.

Tabela 2.2.5. Błędy prognozy godzinnego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym dla hybrydowego modelu neuronowo-heurystycznego

Godz.	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)	Godz.	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
1	5 734	45 024	2,95	17,85	13	9 404	75 128	3,99	24,94
2	5 982	48 629	3,29	32,23	14	9 011	56 653	3,90	25,36
3	5 363	43 410	3,03	29,18	15	8 640	84 081	3,66	31,43
4	4 802	29 749	2,69	20,46	16	10 066	87 559	4,04	33,81
5	4 744	22 744	2,68	14,49	17	11 546	245 102	4,52	69,33
6	5 689	35 994	3,15	18,28	18	9 501	72 582	3,80	26,28
7	9 643	136 834	4,79	70,38	19	9 888	116 248	4,05	39,67
8	9 186	92 392	4,24	44,14	20	9 851	78 784	4,02	33,16
9	7 884	87 711	3,38	37,06	21	9 061	49 750	3,52	23,23
10	7 819	56 593	3,33	27,40	22	7 943	51 960	3,19	21,23
11	7 960	67 791	3,41	27,41	23	7 202	64 635	3,10	29,79
12	8 492	77 793	3,64	27,11	24	6 073	57 685	2,85	28,78
Średnia dla wszystkich godzin						7 978		3,55	
Prognozy dobowego zapotrzebowania na energię						118 366	1 053 697	2,21	15,92

Źródło: W. Bartkiewicz, Z. Gontar, J.S. Zieliński, W. Bardzki, *Neural-heuristic approach to short-term electrical load forecasting problems*, [w:] H. Bothe, R. Rojas (eds), *Neural Computation (NC'2000)*, Berlin 2000, s. 740–744.

Wszystkie dni świąteczne zostały oznaczone w bazie danych systemu i ostateczną prognozę wykonano na bazie analizy typu dnia, na który była ona sporządzana, ze specjalnymi procedurami korekcyjnymi dla wyboru danych wejściowych. Jeżeli, dla przykładu, wejścia wymagane dla modelu dla dnia

specjalnego pochodziły z innego dnia świątecznego (lub z niedzieli), horyzont czasowy prognozy był wydłużany, a baza danych systemu przeszukiwana wstecz aż do znalezienia ostatniego dnia normalnego przed świętem. Procedura ta pozwala uniknąć niepożądanych interakcji pomiędzy wzorcami zapotrzebowania dla różnych typów dni.

Wyniki testowania neuronowo-heurystycznego systemu prognostycznego uwzględniającego wzorce obciążeń dla różnych typów dni przedstawiono w tabeli 2.2.5. Testowanie przeprowadzono na tym samym trzyletnim zbiorze danych, co w przypadku opisanego w poprzednim punkcie modelu MLP dla wszystkich dni (patrz tabela 2.2.2). Porównując błędy w obydwu tabelach, widzimy, że dokładność prognozy uległa znaczącej poprawie. Zarówno maksymalne, jak i średnie błędy dla poszczególnych godzin są obecnie znacznie niższe. Średni błąd prognozy dla wszystkich godzin spadł z poziomu 9907 kWh (4,42%) do 7978 kWh (3,55%), a więc o około 0,9%. Nawet jeszcze bardziej wyrazista jest poprawa dokładności prognoz dobowego zapotrzebowania na energię, z 171 MWh (3,18%) do 118 MWh (2,21%). Kształtuje się więc ona na zbliżonym poziomie jak w przypadku modelu prezentowanego w punkcie 2.2.3. Błąd predykcji jest oczywiście nieznacznie większy, co spowodowane zostało dłuższym, dwudniowym wyprzedzeniem czasowym prognozy.

Zaprezentowany system prognostyczny eksploatowany był również w trybie operacyjnym w spółce dystrybucyjnej. Po kilku miesiącach eksploatacji osiągane wyniki dla zapotrzebowania godzinowego kształtowały się na poziomie około 3,2%. Warto zwrócić uwagę na fakt, że oprócz nietypowych wzorców zapotrzebowania na energię powodowanych przez dni specjalne, na obciążenie sieci mają poważny wpływ również inne nieoczekiwane zdarzenia, takie jak awarie w sieci, duże wydarzenia telewizyjne itp. System prognostyczny umożliwia wprowadzenie przez doświadczonego operatora poprawek do finalnej prognozy. Umożliwiło to redukcję błędów prognozy w tym samym okresie działania modelu do poziomu 2,9% (Bartkiewicz, Gontar, Matusiak, Zieliński i inni 2002; Bartkiewicz, Gontar, Matusiak, Zieliński 2002).

Kolejnym istotnym elementem, który może wpłynąć na poprawę dokładności prognozy zapotrzebowania na energię, są nieregularne cykle produkcyjne dużych odbiorców. Dalsze analizy błędów omawianego systemu prognostycznego wskazały, że w przypadku rozważanej spółki dystrybucyjnej istotnym problemem i źródłem błędów prognozy mogą być różnice w odbiorze energii przez cementownie. Aby zweryfikować poprawność tej hipotezy i zorientować się w skali problemu, przeprowadzono badania porównawcze, przygotowując i testując identyczny system prognostyczny dla tych samych zbiorów danych o obciążeniach, ale z wyłączeniem zakładów cementowniczych (Bartkiewicz, Gontar, Matusiak, Zieliński 2002).

Tabela 2.2.6. Błędy prognozy godzinnego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym po wyłączeniu cementowni

Godz.	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)	Godz.	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
1	3 628	27 227	2,25	18,59	13	5 778	60 134	2,78	35,88
2	3 373	24 225	2,23	19,51	14	6 279	59 623	3,00	34,13
3	3 245	23 173	2,22	19,47	15	6 322	63 760	2,96	34,54
4	3 146	20 242	2,18	17,28	16	6 849	60 981	3,16	35,71
5	3 646	21 281	2,56	19,00	17	7 337	63 329	3,26	23,49
6	3 997	23 731	2,68	21,77	18	7 306	56 423	3,27	23,58
7	5 316	42 882	3,05	29,10	19	7 969	54 471	3,54	31,57
8	5 704	70 575	2,93	41,58	20	6 850	40 945	3,01	28,27
9	5 326	72 431	2,62	40,99	21	5 531	43 842	2,37	20,86
10	5 437	68 985	2,61	37,73	22	4 514	35 699	1,96	15,60
11	5 545	60 684	2,68	35,69	23	4 176	27 751	2,03	13,79
12	5 995	63 015	2,85	37,00	24	3 861	23 623	2,15	15,08
Średnia dla wszystkich godzin						5 297		2,68	
Prognozy dobowego zapotrzebowania na energię						91 569	896 664	1,94	21,28

Źródło: opracowanie własne.

Otrzymane wyniki przedstawione zostały w tabeli 2.2.6.

Porównując wyniki z tabel 2.2.5 i 2.2.6, łatwo możemy zauważyć, że dokładność prognozy godzinnego zapotrzebowania na energię poprawiła się bardzo wyraźnie, z poziomu 3,55% do 2,68%. Największą poprawę zanotowano w godzinach szczytu porannego. Nieregularne cykle produkcyjne dużych odbiorców mają natomiast nieco mniejszy wpływ na dokładność prognozy dobowego zapotrzebowania na energię. Tutaj, jak widzimy, błąd spadł z poziomu 2,21% do 1,94%.

Tego typu czynniki, rzecz jasna, nie mogą zostać uwzględnione na poziomie systemu prognostycznego. Nie są one znane z góry podczas tworzenia prognozy. Przedstawiona analiza wskazuje natomiast, że mają one istotny wpływ na prognozy popytowe, na których opiera się wiele decyzji podejmowanych przez spółki zajmujące się obrotem energią.

Warto więc rozważyć, czy w ramach prowadzonych działań zarządzania stroną popytową (*Demand Side Management*, DSM), czy też zawierania odrębnych umów z dużymi odbiorcami, nie wprowadzić porozumień dotyczących wcześniejszego powiadamiania spółki energetycznej o poważnych zmianach wielkości odbioru.

2.2.6. Prognozy adaptacyjne z wykorzystaniem hybrydowego modelu opartego na sieci MLP i sieci Kohonena

Na koniec przedyskutujemy jeszcze problem wykorzystania sieci neuronowych w procesie prognozy adaptacyjnej, w której model predykcji tworzony (uczony) jest dla konkretnej prognozy. Należy jednak zauważyć, że prezentowane tutaj wyniki mają charakter wstępny. W chwili obecnej podejście adaptacyjne ma bowiem, w opinii autora, dużo większe znaczenie laboratoryjne niż praktyczne. Problemem, jaki napotykamy przy wdrażaniu tego typu systemów, jest konieczność automatyzacji procesu wyboru danych treningowych oraz samego uczenia sieci przy stosunkowo niewielkiej liczbie obserwacji w zbiorze uczącym. Na obecnym etapie przeprowadzenie właściwej selekcji wzorców treningowych wymaga jednak pewnego udziału czynnika ludzkiego – doświadczonego analityka danych. Również uczenie sieci neuronowej dla stosunkowo niewielkich zbiorów danych, z jakimi mamy w tym przypadku do czynienia, jest procesem dosyć złożonym.

Tym niemniej wstępne wyniki laboratoryjne są na tyle obiecujące, że zdecydowaliśmy się przedstawić tę metodę jako możliwą przyszłą gałąź rozwojową technologii prognostycznej w dziedzinie prognozowania zapotrzebowania na energię. Jej idea polega na wyborze zbioru treningowego w postaci obserwacji wzorców zapotrzebowania na energię najbardziej podobnych do wzorca wejściowego danych dla konkretnej prognozy. Do wstępnej analizy bazy danych historycznych obciążeń i wyboru zbioru obserwacji treningowych w naszych badaniach wykorzystaliśmy sieć neuronową SOM (*Self Organising Map*) Kohonena (Bardzki, Bartkiewicz 1995).

Sieci SOM Kohonena stanowią kolejną standardową architekturę sieci neuronowych wykorzystywanych głównie w procesie analizy, grupowania i wizualizacji danych. Szczegóły ich działania znaleźć można w każdej niemal pozycji z literatury przedmiotu (patrz dla przykładu Zieliński 2000; Hertz, Krogh, Palmer 1993; Korbicz, Obuchowicz, Uciński 1994; Żurada, Barski, Jędruch 1996). Składają się one zwykle z warstwy wejściowej oraz z jednej warstwy neuronów przetwarzających, zwanej warstwą konkurencyjną lub Kohonena.

Jak wspomnieliśmy, sieci Kohonena wiążą się z procesami grupowania i klasyfikacji danych. Uformowane w procesie uczenia kategorie reprezentowane są przez neurony warstwy konkurencyjnej. Dane podobne, które powinny być zakwalifikowane do tej samej grupy, aktywizują ten sam neuron. Jego wagi tworzą tzw. centroid klasy w wielowymiarowej przestrzeni wejść.

Podstawową operacją wykonywaną przez neurony w warstwie konkurencyjnej sieci Kohonena jest wyznaczanie odległości między wagami neuronu $\mathbf{w}$ a wektorem wejściowym sieci $\mathbf{x}$. Wyjście każdego i-tego neuronu y_i obliczane jest więc jako (Zieliński 2000):

$$y_i = \left\| \mathbf{x} - \mathbf{w}_i \right\| \tag{2.2.16a}$$

Zwykle stosowana jest metryka euklidesowa:

$$\mathrm{y}_i = \sqrt{\sum_{j=1}^{n} (x_j - w_{ij})^2} \tag{2.2.16b}$$

Proces uczenia sieci SOM, w odróżnieniu od omawianego wcześniej algorytmu wstecznej propagacji błędu, ma charakter nienadzorowany. Dane treningowe nie zawierają żadnej informacji na temat pożądanych wyjść. Sieć ma za zadanie samodzielnie pogrupować niezaetykietowane dane, jedynie na podstawie występujących w nich korelacji. Algorytm uczenia stosowany w sieciach Kohonena polega na tzw. uczeniu konkurencyjnym lub uczeniu typu „zwycięzca bierze wszystko". Neurony sieci, w odpowiedzi na sygnał wejściowy, rywalizują ze sobą. Rywalizację tę wygrywa jeden neuron, na ogół najsłabiej reagujący. Zwycięski neuron oraz ewentualnie jego najbliższe otoczenie podlegają procesowi uczenia polegającemu na zbliżeniu (w różnym stopniu) ich wag do wektora wejściowego.

Proces uczenia rozpoczynamy więc od (zazwyczaj losowej) inicjacji wag. Następnie iteracyjnie prezentujemy kolejne wektory ze zbioru treningowego. Dla każdego z nich znajdujemy najsłabiej reagujący neuron w warstwie konkurencyjnej (Zieliński 2000):

$$y_c = \min_{1 \le i \le M} (y_i) = \min_{1 \le i \le M} (\| \mathbf{x} - \mathbf{w}_i \|) \tag{2.2.17}$$

gdzie M jest liczbą neuronów sieci.

Następnie wektor wag każdego z neuronów sieci $\mathbf{w}_i$ modyfikowany jest zgodnie z regułą:

$$\Delta w_{ij} = \eta h(c,i)(x_j - w_{ij}) \tag{2.2.18}$$

gdzie:

η – współczynnik uczenia,

c – indeks neuronu zwycięskiego,

$h(c, i)$ – funkcja sąsiedztwa,

$1 \le i \le M$ – indeks neuronu; M – liczba neuronów w warstwie konkurencyjnej,

$1 \le j \le n$ – współrzędna wektora wagowego; n – liczba wejść sieci.

Algorytm uczenia sieci SOM Kohonena działa więc na zasadzie korekty wag zwycięskiego neuronu tak, by „przesunąć" je, w pewnym stopniu, w stronę prezentowanego sieci wzorca danych wejściowych. W ten sposób, po wielokrotnej prezentacji zbioru treningowego, wektory wag neuronów sieci stają się uśrednionymi centrami skupień wzorców danych, dla których zwycięża dany neuron.

W każdym kroku modyfikowane są nie tylko wagi neuronu zwycięskiego, ale również innych węzłów sieci SOM. Wagi jednostek położonych blisko zwycięzcy są modyfikowane silniej niż tych bardziej odległych. Dzięki temu uzyskuje się przydatny w wielu zastosowaniach efekt płynnego przejścia pomiędzy skupieniami. Skupiska reprezentowane przez neurony położone bliżej siebie w strukturze sieci grupują wzorce danych bardziej podobne do siebie niż w przypadku neuronów odległych. Efekt ten osiągany jest poprzez zastosowanie w procesie uczenia (2.2.18) funkcji sąsiedztwa regulującej stopień modyfikacji wag dla danego neuronu. Funkcja ta powinna spełniać następujące warunki:

– $h(c, c) = 1$,

– $h(c, i) \to 0$, w miarę jak rośnie odległość i-tego neuronu oraz zwycięskiego c, w sensie ich położenia w strukturze sieci,

– współczynnik uczenia η oraz sąsiedztwo powinny maleć w miarę postępów uczenia.

Klasy, jak już wspomnieliśmy, formowane są przez wektory wejściowe, dla których zwycięża ten sam neuron. Mają one formę kul w wielowymiarowej przestrzeni wejść, których środek (zwany centroidem lub prototypem) wyznaczony jest przez wektor wagowy odpowiadającego im neuronu. Proces adaptacyjnego wyboru danych do uczenia modelu prognostycznego polega więc na:

– treningu sieci SOM Kohonena dla wzorców odpowiadających wejściom prognozy, przygotowanym z bazy danych historycznych obserwacji obciążeń sieci elektroenergetycznej i warunków atmosferycznych; w wyniku otrzymujemy wektory wagowe neuronów sieci Kohonena odpowiadające poszczególnym grupom (skupieniom) analizowanych wzorców,

– uformowaniu samych skupień; wszystkie wzorce prezentowane są ponownie sieci Kohonena; wzorce, dla których „wygrywa" ten sam neuron, należą do tego samego skupienia,

– dla konkretnej prognozy znajdowany jest neuron „zwycięski", czyli najsłabiej reagujący dla zestawu danych wejściowych tej prognozy; wzorce historyczne należące do skupienia reprezentowanego przez ten neuron posłużą do budowy (treningu) właściwego modelu prognostycznego.

Omawianą metodę zastosowano do budowy modelu adaptacyjnego prognozy szczytowego zapotrzebowania na energię w szczycie wieczornym (Bardzki, Bartkiewicz 1995):

$$ZSW_d = f(ZSR_{d-1}, ZSW_{d-1}, TR_{d-1}, TP_{d-1}, TW_{d-1}, ZSR_d, TR_d, TP_d, TW_d) \quad (2.2.19)$$

gdzie:

ZSW_d – zapotrzebowanie na energię w szczycie wieczornym, w dniu d,
ZSR_d – zapotrzebowanie na energię w szczycie porannym, w dniu d,
TR_d – temperatura poranna (mierzona o godzinie 8), w dniu d,
TP_d – temperatura w południe (mierzona o godzinie 13), w dniu d,
TW_d – temperatura wieczorna (mierzona o godzinie 21), w dniu d.

Do samej prognozy zapotrzebowania szczytowego (2.2.19) wykorzystano model warstwowej sieci perceptronowej MLP, o strukturze {9, 9, 1} (z dziewięcioma neuronami w warstwie ukrytej). Wyniki wstępnego testowania modelu znajdują się w tabeli 2.2.7. Zawiera ona porównanie dokładności modelu nieadaptacyjnego (model 1) oraz adaptacyjnego, w którym sieć uczona była do każdej prognozy, zgodnie z procedurą opisaną w obecnym punkcie, z wykorzystaniem sieci Kohonena (model 2). Jak widzimy, podejście adaptacyjne pozwoliło na redukcję błędu mniej więcej o połowę.

Tabela 2.2.7. Błędy prognozy adaptacyjnej z wykorzystaniem hybrydowego modelu opartego na sieci MLP i sieci Kohonena

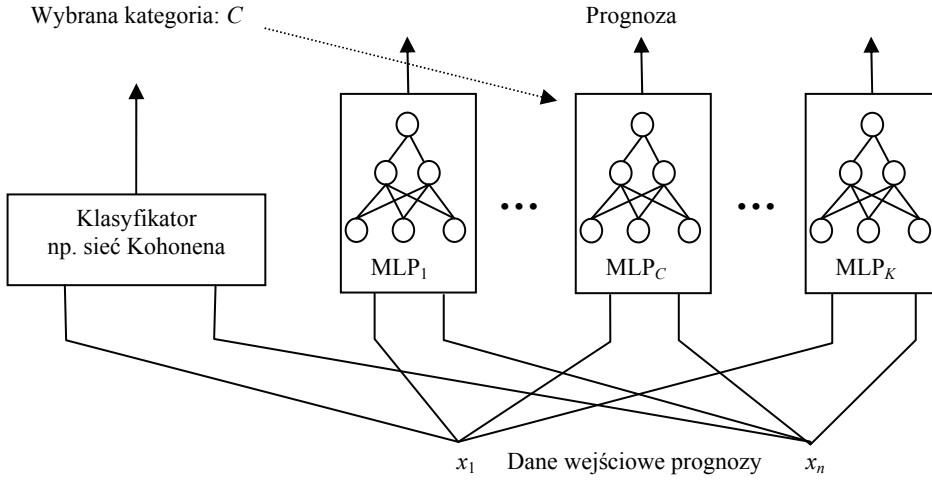
Dzień	Rzeczywiste ZSW	Model 1		Model 2	
	(MW)	Błąd (MW)	%	Błąd (MW)	%
4.09.1992	488	32,7	6,71	4,7	0,96
5.09.1992	552	35,8	6,50	14,8	2,68
6.09.1992	484	28,3	5,85	7,7	1,58
7.09.1992	547	1,4	0,26	8,6	1,56
11.09.1992	515	13,4	2,60	13,9	2,70
12.09.1992	478	27,9	5,84	6,8	1,41
13.09.1992	514	7,8	1,53	2,6	0,50
14.09. 1992	487	18,3	3,76	2,8	0,57
18.09.1992	537	50,0	9,47	1,6	0,30
19.09.1992	496	25,0	5,04	9,3	1,87
20.09.1992	542	0,1	0,02	6,3	1,17
21.09.1992	509	7,2	1,42	7,2	1,42
25.09.1992	524	6,7	1,27	9,0	1,70
26.09.1992	560	4,8	7,28	25,2	4,49
27.09.1992	555	3,6	0,65	15,7	2,82
28.09.1992	541	16,1	2,99	3,9	0,72
Średnie błędy		19,69	3,82	8,76	1,65
Błędy maksymalne		50	9,47	25,2	4,49

Źródło: opracowanie własne.

Należy jednak pamiętać, że przedstawione w tabeli 2.2.7 wyniki mają charakter wstępny i prezentują raczej to, co można uzyskać za pomocą podejścia adaptacyjnego. Metoda ta ma szereg parametrów, których dobór wymaga jednak udziału czynnika ludzkiego (analityka danych). Jak do tej pory, w opisywanym przypadku nie udało się opracować obiektywnej procedury algorytmicznej, możliwej do implementacji maszynowej.

2.2.7. Prognozy zapotrzebowania na energię z wykorzystaniem lokalnych modeli MLP

Rozwiązaniem pośrednim pomiędzy zastosowaniem globalnego modelu prognostycznego, wykorzystywanego do wszystkich prognoz, oraz modelu adaptacyjnego, tworzonego do konkretnej prognozy, może być wcześniejsze przygotowanie grupy modeli lokalnych skonstruowanych dla jednorodnych podsegmentów danych, a następnie wybór właściwego modelu dla danej prognozy. Pozwala to na przejście od uczenia globalnego sieci w całej dziedzinie aproksymowanego odwzorowania do lokalnego otoczenia wektora danych wejściowych dla konkretnej prognozy.



Rysunek 2.2.6. Przełączanie lokalnych modeli MLP przy użyciu klasyfikatora
Źródło: opracowanie własne

Z drugiej strony, model nie musi być uczony do każdej prognozy, tak jak to przyjmowaliśmy w punkcie 2.2.6; odpowiednie prace nad jego przygotowaniem

mogą być wykonane z góry przez analityka danych. Ponadto dane dzielone są zazwyczaj jedynie na kilka segmentów, które zawierają znacznie większą liczbę wzorców treningowych, przez co proces uczenia sieci nie jest tak czuły na występujące w nich anomalie.

Jednorodne segmenty danych otrzymywane są przy zastosowaniu algorytmu grupowania danych. Wybór metody wykorzystywanej w tym procesie nie jest istotny. Może być to, dla przykładu, zastosowana w poprzednim podpunkcie sieć SOM Kohonena. Dane należące do każdego segmentu wykorzystywane są do uczenia odrębnej sieci MLP.

Prognoza ma charakter dwukrokowy. W pierwszym kroku określony zostaje segment, do którego należy wzorzec danych wejściowych. W przypadku sieci Kohonena określany jest on przez neuron reagujący najsłabiej dla tego wzorca. Końcową prognozę (drugi krok) otrzymuje się przy wykorzystaniu sieci MLP uzyskanej dla odpowiedniego segmentu (rysunek 2.2.6).

Zaprezentowany model zastosowany został do prognozy szczytowego zapotrzebowania na energię (maksymalnego godzinnego zapotrzebowania na energię) z dwudniowym wyprzedzeniem czasowym (Bartkiewicz 1998a):

$$ZS_d = f(ZS_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZS(t)_{d-2}, ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \tag{2.2.20}$$

gdzie oznaczenia są analogiczne jak w poprzednich modelach:

$ZS(t)_d$ – zapotrzebowanie na energię w szczytowej godzinie, w dniu d,

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d, gdzie t oznacza godzinę szczytową,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Model prognostyczny przetestowano dla półrocznego okresu obejmującego wszystkie dni tygodnia (włączając niedziele), ale nieuwzględniającego dni specjalnych (patrz punkt 2.2.5). Wykonano również badania porównawcze dla prognoz uzyskanych za pomocą pojedynczej sieci neuronowej MLP. Otrzymane wyniki zaprezentowane zostały w tabeli 2.2.8. Jak widzimy, zastosowanie zespołu kilku lokalnych modeli MLP pozwoliło w tym przypadku na nieznaczną poprawę dokładności prognozy.

Tabela 2.2.8. Błędy prognozy szczytowego godzinnego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym

Model	MAE (kWh)	MAX AE (kWh)	RMSE (kWh)	MAPE (%)	MAX APE (%)
Sieć MLP	10 715	55 755	14 288	4,65	31,26
Zespół lokalnych MLP	10 236	55 340	14 131	4,35	31,79

Źródło: opracowanie własne.

Do zbliżonych rozwiązań, wykorzystujących podejścia neuronowo-rozmyte, będziemy jeszcze wracać w punktach 2.3.3–2.3.5.

2.3. Prognozowanie zapotrzebowania na energię z wykorzystaniem modeli neuronowo-rozmytych

2.3.1. Lingwistyczne systemy z logiką rozmytą (MISO)

Pojęcie zbioru rozmytego wprowadził w 1965 r. Lotfi A. Zadeh (Zadeh 1965). Niech X będzie pewną przestrzenią rozważanych obiektów. Zbiór rozmyty A definiowany jest przez parę:

$$\{X, \mu_A\} \tag{2.3.1}$$

gdzie μ_A: $X \rightarrow [0, 1]$ jest funkcją, która dla każdego elementu z X określa, w jakim stopniu przynależy on do zbioru A. Funkcję μ_A nazywamy funkcją przynależności zbioru A. W przypadku zbioru rozmytego mamy więc płynne przejście między całkowitą przynależnością ($\mu_A(x) = 1$) a nieprzynależnością ($\mu_A(x) = 0$).

Naturalnie pełniejsze omówienie zarówno teorii zbiorów rozmytych, jak i jej wykorzystania w systemach z logiką rozmytą, jest zbyt obszernym zagadnieniem, aby mogło zostać przedstawione na łamach tej książki. Zainteresowanych Czytelników odsyłamy do szczegółowych pozycji z zakresu literatury przedmiotu, dla przykładu: Rutkowska, Piliński, Rutkowski 1997; Yager, Filev 1995; Zieliński 2000. W bieżącym punkcie przedstawimy jedynie wprowadzenie do pewnej podklasy systemów z logiką rozmytą, tzw. addytywnych systemów rozmytych z regułą wnioskowania Larsena stanowiących podstawę większości wykorzystywanych dalej architektur neuronowo-rozmytych. Zakładać przy tym będziemy, że model ma wiele wejść i jedno wyjście (*Multi Input Single Output*, MISO), a ponadto zarówno wejścia, jak i wyjścia mają charakter liczbowy.

Podstawowy element systemu rozmytego stanowi zdanie postaci:

$$V \text{ jest } A \tag{2.3.2}$$

Uznajmy, że zmienna V przyjmuje wartości w pewnej przestrzeni X. Nie są one przy tym określone precyzyjnie, a nieprecyzja ta modelowana jest za pomocą zbiorów rozmytych. Zdanie (2.3.2) oznacza więc, że zmienna V przyjmuje wartości ograniczane przez pewien zbiór rozmyty A zdefiniowany w przestrzeni X. Zmienną V nazywamy wówczas zmienną lingwistyczną.

Funkcje przynależności zbiorów rozmytych mogą mieć dowolną postać, ze względów praktycznych zwykle przyjmuje się jednak jedną z trzech form: trójkątną, trapezoidalną lub gaussowską. Wybór rodzaju reprezentacji zbiorów rozmytych zależy od konkretnego przypadku i powinien być wykonany w czasie analizy rozważanego problemu. Podział przestrzeni X na zbiory trójkątne jest bardzo popularny, zwłaszcza w zastosowaniach inżynieryjnych, związanych ze sterowaniem. W zastosowaniach do celów zarządzania sugeruje się raczej wykorzystanie krzywych gaussowskich.

W systemie MISO, $y(x_1, \ldots, x_n)$: $R^n \rightarrow R$ zależności między zmiennymi lingwistycznymi opisywane są zwykle za pomocą zbioru reguł postaci (Jabłoński, Bartkiewicz 2006; Zieliński 2000):

$$\begin{aligned} &\textbf{JEŻELI } X_1 \text{ jest } A_{11} \textbf{ I } \ \ldots \ \textbf{ I } X_n \text{ jest } A_{1n} \quad \textbf{TO } Y \text{ jest } B_1 \\ &\textbf{JEŻELI } X_1 \text{ jest } A_{K1} \textbf{ I } \ \ldots \ \textbf{ I } X_n \text{ jest } A_{Kn} \quad \textbf{TO } Y \text{ jest } B_K , \end{aligned} \tag{2.3.3}$$

gdzie A_{ij} są zbiorami rozmytymi definiującymi wartości zmiennej wejściowej X_j, $j = 1,\ldots, n$ dla każdej i-tej reguły, $i = 1, \ldots, K$, zaś B_i są zbiorami rozmytymi definiującymi dla tych reguł wartości zmiennej wyjściowej Y.

Dla konkretnego wzorca wartości wejściowych systemu $x_1, \ldots, x_n$ wynikiem działania każdej z reguł jest zbiór rozmyty B'_i, którego funkcję przynależności możemy zdefiniować następująco:

$$\mu_{B'_i}(y) = \tau_i \mu_{B_i}(y), \qquad \tau_i = \prod_{j=1}^{n} \mu_{A_{ij}}(x_j), \quad i = 1,\ldots,K \tag{2.3.4}$$

Wielkość τ_i nazywamy stopniem prawdziwości i-tej reguły dla danego wzorca wejściowego $x_1, \ldots, x_n$. Wykorzystana w (2.3.4) reguła wnioskowania w systemie rozmytym, w której zastosowano iloczyn algebraiczny, nazywana jest regułą Larsena. Nie jest to jedyny możliwy wybór; do kojarzenia poprzednika i następnika reguł możemy wykorzystać dowolną operację normy trójkątnej. Popularne podejście stanowi np. tzw. reguła Mamdaniego, w której wykorzystu-

je się w (2.3.4), zamiast iloczynu, operację minimum. Jednakże w systemach neuronowo-rozmytych uczonych na podstawie danych reguła Larsena jest podejściem wygodniejszym, choćby z powodu różniczkowalności iloczynu.

Kolejnym krokiem w systemie rozmytym musi być scalenie wyników działania poszczególnych reguł. W wyniku tej operacji, na podstawie zbiorów B'_i, $i = 1, \ldots, K$, otrzymujemy jeden zbiór rozmyty B' będący łącznym wynikiem działania wszystkich reguł systemu dla danego wejścia $x_1, \ldots, x_n$. W naszym przypadku wykorzystamy sumę ważoną:

$$\mu_{B'}(y) = \sum_{i=1}^{K} w_i \mu_{B'_i}(y) \tag{2.3.5}$$

gdzie w_i, $i =, 1, \ldots, K$ są wagami dobieranymi do konkretnego przypadku tak, by wartości funkcji przynależności znajdowały się w przedziale [0, 1]. Podobnie jak w przypadku reguły wnioskowania, w systemach rozmytych stosować można wiele rozmaitych metod scalania wyników. Systemy, w których wykorzystuje się (2.3.5), nazywane są addytywnymi systemami rozmytymi.

Jak już wspomnieliśmy, interesuje nas reprezentacja wyniku działania systemu w postaci liczbowej. Ostatnim krokiem musi więc być wyostrzenie (defuzyfikacja) wynikowego zbioru rozmytego B', tzn. znalezienie takiej wartości liczbowej, która w jak najlepszy sposób reprezentować będzie cały zbiór rozmyty. Defuzyfikacja przeprowadzana jest zwykle za pomocą jednej z dwu podstawowych metod: metody środka obszaru (*Center of Area*, COA), zwanej również metodą centroidu, lub drugiej metody, średniej maksymalnej (*Mean of Maximum*, MOM). Pierwszą metodę stosuje się w przypadku systemów rozmytych modelujących ciągłe zależności między wejściem a wyjściem, drugą – zazwyczaj w przypadku zależności skokowych.

Metoda centroidu wyznacza niejako „środek ciężkości" zbioru rozmytego:

$$y^* = \frac{\int_Y y \mu_{B'}(y) dy}{\int_Y \mu_{B'}(y) dy} \tag{2.3.6}$$

Podsumowując, na podstawie (2.3.4)–(2.3.6), w addytywnych systemach rozmytych MISO z regułą wnioskowania Larsena, wyjście systemu $y(x_1, \ldots, x_n)$ dla danego wzorca wejściowego $x_1, \ldots, x_n$ wyznaczyć można zgodnie z zależnością:

$$y(x_1,\ldots,x_n) = \frac{\int_Y y \sum_{i=1}^{K} w_i \mu_{B'_i}(y) dy}{\int_Y \sum_{i=1}^{K} w_i \mu_{B'_i}(y) dy} = \frac{\int_Y y \sum_{i=1}^{K} w_i \tau_i \mu_{B_i}(y) dy}{\int_Y \sum_{i=1}^{K} w_i \tau_i \mu_{B_i}(y) dy}, \quad \tau_i = \prod_{j=1}^{n} \mu_{A_{ij}}(x_j) \tag{2.3.7}$$

Jak więc widzimy, procedura wnioskowania w systemie z logiką rozmytą ma charakter w zasadzie czysto ilościowy i numeryczny. Z drugiej jednak strony, struktura systemu i jego elementy mają charakter logiczno-symboliczny, możliwy do interpretacji i analizy. Tłumaczy to w dużej mierze popularność rozwiązań opartych na podejściu neuronowo-rozmytym, łączących możliwości uczenia maszynowego z interpretowalnością uzyskanej wiedzy (Lin, Lee 1996; Rutkowska 1997; Jang, Sun, Mizutani 1997; Zieliński 2000).

2.3.2. Prognozowanie zapotrzebowania na energię z wykorzystaniem sieci neuronowo-rozmytych typu FBF

Podstawę działania modeli neuronowo-rozmytych typu FBF (*Fuzzy Basis Function*) stanowią opisane w poprzednim punkcie addytywne systemy z logiką rozmytą, z regułą wnioskowania Larsena. Na podstawie równania systemu (2.3.7) możemy zapisać:

$$y(x_1,...,x_n)=\frac{\int_Y y\sum_{i=1}^{K} w_i\tau_i\mu_{B_i}(y)dy}{\int_Y \sum_{i=1}^{K} w_i\tau_i\mu_{B_i}(y)dy}=\frac{\sum_{i=1}^{K} w_i\tau_i\int_Y y\mu_{B_i}(y)dy}{\sum_{i=1}^{K} w_i\tau_i\int_Y \mu_{B_i}(y)dy} \tag{2.3.8}$$

Oznaczmy przez b_i^*, $i = 1, \ldots, K$ wyostrzone (zdefuzyfikowane) środki zbiorów rozmytych B_i występujących w następnikach reguł systemu, tj. na podstawie (2.3.6) $b_i^*=\frac{\int_Y y\mu_{B_i}(y)dy}{\int_Y \mu_{B_i}(y)dy}$. Wówczas zależność (2.3.8) możemy zapisać następująco:

$$y(x_1,...,x_n)=\frac{\sum_{i=1}^{K} w_i\tau_i b_i^*\int_Y \mu_{B_i}(y)dy}{\sum_{i=1}^{K} w_i\tau_i\int_Y \mu_{B_i}(y)dy} \tag{2.3.9}$$

Przyjmując dalej jako wartości wag $w_i=\frac{1}{\int_Y \mu_{B_i}(y)dy}$, otrzymujemy tzw. uproszczoną metodę wnioskowania rozmytego (Yager, Filev 1995):

$$y(x_1,...,x_n)=\frac{\sum_{i=1}^{K} \tau_i b_i^*}{\sum_{i=1}^{K} \tau_i} \tag{2.3.10}$$

Przyjęte wartości w_i są dopuszczalne, ponieważ nie zależą one od konkretnego wejścia systemu x_1, …, x_n, a więc są stałe. Zauważmy również, że metoda wnioskowania (2.3.10) rzeczywiście ma charakter uproszczony, ponieważ wynik działania systemu zależy wyłącznie od poziomu prawdziwości reguł dla danego wejścia oraz od środków zbiorów rozmytych następników. Taki sposób definicji wag powoduje, że następniki reguł przyjmują charakter liczbowy. Rezygnujemy z dodatkowej informacji niesionej przez całość funkcji przynależności zbiorów B_i.

Model sieci neuronowo-rozmytej FBF przyjmuje zatem uproszczoną metodę wnioskowania rozmytego (Wang, Mendel 1992). Zakłada się ponadto, że funkcje przynależności zbiorów rozmytych A_{ij}, j =1,…, n, i = 1, …, K w poprzednikach reguł mają charakter funkcji Gaussa, tzn. na podstawie (2.3.7) możemy napisać:

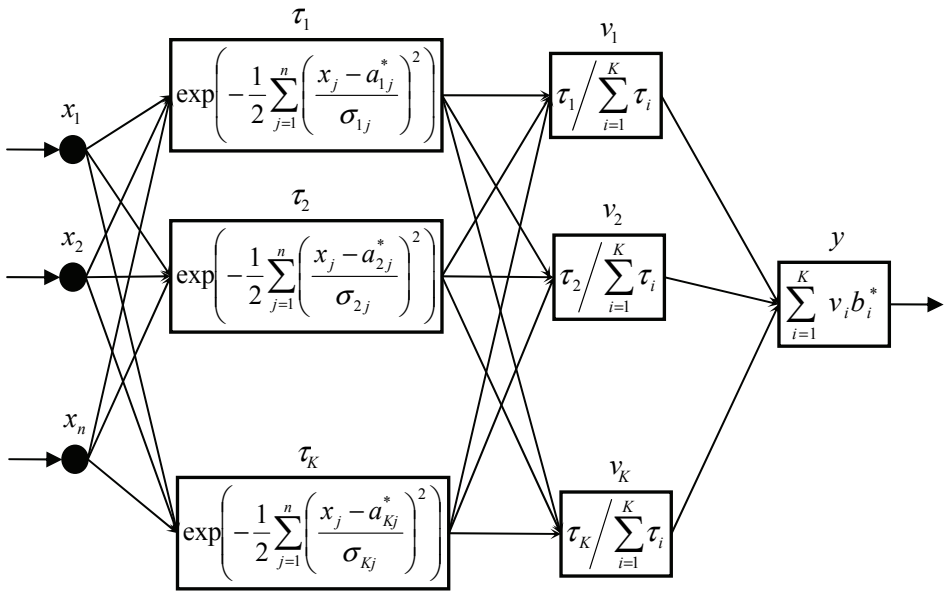
$$\tau_i = \prod_{j=1}^{n} \mu_{A_{ij}}(x_j) = \prod_{j=1}^{n} \exp\left(\frac{-(x_j - a_{ij}^*)^2}{2\sigma_{ij}^2}\right) = \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^2}\right) \qquad (2.3.11)$$

gdzie a_{ij}^* i σ_{ij}, j = 1,…, n, i = 1, …, K są parametrami krzywej Gaussa, odpowiednio parametrem definiującym środek (centroid) i szerokość funkcji przynależności zbioru rozmytego A_{ij}. Ostatecznie więc na podstawie zależności (2.3.10) i (2.3.11) równanie sieci FBF możemy zapisać następująco:

$$y(x_1,...,x_n) = \frac{\sum_{i=1}^{K} b_i^* \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^2}\right)}{\sum_{i=1}^{K} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^2}\right)} \qquad (2.3.12)$$

Model rozmyty FBF (2.3.12) można przedstawić w postaci jednokierunkowej sieci neuronowej zaprezentowanej na rysunku 2.3.1. Sieć składa się z dwu warstw ukrytych neuronów. Pierwsza z nich zawiera K jednostek, których wyjściem są stopnie prawdziwości poprzedników reguł τ_i. Parametrami adaptacyjnymi (wagami) neuronów tej warstwy są a_{ij}^* i σ_{ij}, j = 1,…, n, i = 1, …, K, czyli centroidy i szerokości funkcji przynależności zbiorów rozmytych poprzedników reguł A_{ij}.

Druga warstwa ukryta nie posiada wag i wykonuje jedynie normalizację prawdziwości poprzedników reguł. Zawiera również K neuronów, których wyjścia oznaczone są na rysunku 2.3.1 przez v_i. Ostatnia z warstw, wyjściowa, zawiera jeden neuron, którego wyjście y stanowi jednocześnie wyjście całej sieci i obliczane jest jako iloczyn skalarny v_i oraz centroidów zbiorów rozmytych następników reguł b_i^*, które również wchodzą w skład adaptacyjnych wag sieci.



Rysunek 2.3.1. Struktura sieci neuronowo-rozmytej FBF
Źródło: R.R.Yager, D.P. Filev, *Podstawy modelowania i sterowania rozmytego*, Warszawa 1995

Jak więc widzimy, zbiór wag sieci FBF tworzony jest przez odpowiednie parametry zbiorów rozmytych poprzedników i następników reguł, a_{ij}^*, σ_{ij} oraz b_i^* j =1,…, n, i = 1, …, K. Ponieważ, w przeciwieństwie do np. sieci MLP, wagi te mają swoją interpretację fizyczną, można do ich wyznaczenia wykorzystać różnego rodzaju specyficzne metody znajdowania funkcji przynależności zbiorów rozmytych, np. oparte na analizie skupień w danych. Tę właściwość systemów neuronowo-rozmytych będziemy jeszcze wykorzystywać w kolejnych modelach prezentowanych w bieżącym rozdziale. Możliwe jest jednak również uczenie sieci FBF metodą najmniejszych kwadratów, zbliżoną do algorytmu wstecznej propagacji błędu opisanego w punkcie 2.2.2 (Wang, Mendel 1992; Jang, Sun, Mizutani 1997; Yager, Filev 1995), które to podejście zastosowaliśmy do bieżącego zagadnienia.

Uczenie ma, oczywiście, charakter nadzorowany, a więc zbiór treningowy składa się ze wzorców wejściowych oraz odpowiadających im znanych (treningowych) wartości zmiennej wyjściowej $\{\mathbf{x}_k, t_k\} = \{(x_{k1}, \ldots, x_{kn}), t_k\}$, $k = 1, \ldots, N$. Cel uczenia polega na minimalizacji błędu kwadratowego pomiędzy wartościami treningowymi a faktycznymi wyjściami sieci:

$$E = \frac{1}{2}\sum_k (t_k - y(x_{k1},...,x_{kn}))^2 = \frac{1}{2}\sum_k (t_k - y_k)^2 \qquad (2.3.13)$$

Dla sieci FBF możemy stosunkowo łatwo wyznaczyć gradient pochodnych błędu E względem parametrów wagowych (patrz załącznik 2, punkt 1):

$$\frac{\partial E}{\partial b_i^*} = (y_k - t_k)v_i$$

$$\frac{\partial E}{\partial a_{ij}^*} = (y_k - t_k)v_i \frac{x_i - a_{ij}^*}{\sigma_{ij}^2}\left(b_i^* - y_k\right) \tag{2.3.14}$$

$$\frac{\partial E}{\partial \sigma_{ij}} = (y_k - t_k)v_i \frac{(x_i - a_{ij}^*)^2}{\sigma_{ij}^3}\left(b_i^* - y_k\right) \qquad j = 1,\ldots, n,\ i = 1, \ldots, K$$

Uczenie, podobnie jak w metodzie wstecznej propagacji błędu, polega na wielokrotnej prezentacji sieci zbioru uczącego i wyznaczaniu dla każdego ze wzorców treningowych poprawki dla wag:

$$\Delta b_i^* = -\eta \frac{\partial E}{\partial b_i^*}$$

$$\Delta a_{ij}^* = -\eta \frac{\partial E}{\partial a_{ij}^*} \tag{2.3.15}$$

$$\Delta \sigma_{ij} = -\eta \frac{\partial E}{\partial \sigma_{ij}}$$

Sieć FBF zastosowana została do trzech problemów związanych z krótkoterminowymi prognozami zapotrzebowania na energię elektryczną.

Model 1

Model prognozy dobowego zapotrzebowania na energię elektryczną z jednodniowym wyprzedzeniem czasowym. Prognoza wykonywana jest na podstawie wzorca zapotrzebowania w dniu poprzednim oraz informacji o zmianie temperatur.

$$\begin{aligned} ZD_d = f(&ZG_{d-1}(1),...,ZG_{d-1}(24), TMIN_{d-1}, \\ &TMAX_{d-1}, TMIN_d, TMAX_d, dt_{1d},...,dt_{6d}) \end{aligned} \tag{2.3.16}$$

gdzie:

ZD_d – dobowe zapotrzebowanie na energię w dniu d,

$ZG_{d-1}(1)$, ..., $ZG_{d-1}(24)$ – godzinowy rozkład zużycia w dniu poprzednim,

$TMIN_{d-1}$, $TMAX_{d-1}$ – temperatura minimalna i maksymalna w dniu poprzednim,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Model 1 ma identyczną strukturę jak model prognozy zapotrzebowania na energię (2.2.10) omawiany w punkcie 2.2.3, w którym wykorzystano sieć MLP. Sieć neuronowo-rozmyta FBF zastosowana została do predykcji obciążeń w tym samym systemie elektroenergetycznym i przetestowana dla tych samych danych. W tabeli 2.3.1 przedstawiono porównanie dokładności działania sieci FBF z prezentowanymi wcześniej w tabeli 2.2.1 wynikami otrzymanymi przy użyciu sieci MLP i regresji liniowej (Bartkiewicz 1998b, c; Bartkiewicz, Zieliński 1998; Bartkiewicz, Butkevych i inni 2001).

Dokładność prognozy osiągnięta przez system neuronowo-rozmyty FBF kształtowała się na poziomie zbliżonym do tej uzyskanej za pomocą sieci neuronowej. Otrzymany średni błąd procentowy MAPE był nieco niższy i wyniósł poniżej 2%. Wyraźnie jednak wzrósł błąd maksymalny, co zaowocowało również nieco wyższą wartością błędu kwadratowego. Dokładność prognoz uzyskanych przez modele nieliniowe była jednak wyraźnie lepsza niż w przypadku regresji liniowej.

Model 2

Sieć FBF zastosowano do takiego samego problemu prognozy dobowego zapotrzebowania na energię z jednodniowym wyprzedzeniem czasowym jak w modelu 1, ale dla innej spółki dystrybucyjnej. Struktura modelu jest identyczna jak określona w (2.3.16).

Tabela 2.3.1. Błędy prognozy dobowego zapotrzebowania na energię z jednodniowym wyprzedzeniem czasowym (model 1)

Model	MAE (MW)	MAX AE (MW)	RMSE (MW)	MAPE (%)	MAX APE (%)
Regresja liniowa	119	953	176	2,31	22,25
Sieć MLP	106	470	141	2,04	10,83
Sieć FBF	102	630	148	1,99	14,51

Źródło: opracowanie własne.

Tabela 2.3.2. Błędy prognozy dobowego zapotrzebowania na energię z jednodniowym wyprzedzeniem czasowym (model 2)

Model	MAE (MW)	MAX AE (MW)	RMSE (MW)	MAPE (%)	MAX APE (%)
Regresja liniowa	130	758	190	2,21	17,07
Sieć MLP	95	725	149	1,62	16,34
Sieć FBF	94	705	147	1,60	15,89

Źródło: opracowanie własne.

W tabeli 2.3.2 przedstawiono wyniki działania sieci FBF (Bartkiewicz 1998c). Również i w tym przypadku przeprowadzono badania porównawcze z prognozami uzyskanymi dla tego samego systemu elektroenergetycznego za pomocą modelu neuronowego MLP i regresji liniowej. Jak widzimy, także dla modelu 2 nieliniowe sieci neuronowe MLP i neuronowo-rozmyte FBF uzyskały porównywalną dokładność, wyraźnie lepszą niż klasyczny model liniowy.

Model 3

Trzecim problemem, do którego zastosowano sieć neuronowo-rozmytą FBF, jest prognoza mocy szczytowej, w szczycie wieczornym, z półdniowym wyprzedzeniem czasowym. Przyjęto model o następującej strukturze:

$$ZSW_d = f(ZSR_{d-1}, ZSW_{d-1}, TR_{d-1}, TP_{d-1}, TW_{d-1}, ZSR_d, TR_d, TP_d, TW_d) \tag{2.3.17}$$

gdzie:

ZSW_d – zapotrzebowanie na energię w szczycie wieczornym, w dniu d,
ZSR_d – zapotrzebowanie na energię w szczycie porannym, w dniu d,
TR_d – temperatura poranna (mierzona o godzinie 8), w dniu d,
TP_d – temperatura w południe (mierzona o godzinie 13), w dniu d,
TW_d – temperatura wieczorna (mierzona o godzinie 21), w dniu d.

Również i w tym przypadku oprócz zastosowania sieci FBF wykonano modele prognostyczne oparte na sieciach neuronowych MLP i regresji liniowej. Porównanie otrzymanych wyników znajduje się w tabeli 2.3.3 (Bartkiewicz 1998c). Podobnie jak w poprzednich przypadkach możemy mówić o porównywalnej dokładności prognoz uzyskanych za pomocą sieci FBN i MLP, natomiast o wyraźnie gorszym działaniu modelu liniowego.

Tabela 2.3.3. Błędy prognozy szczytowego zapotrzebowania z półdniowym wyprzedzeniem czasowym (model 3)

Model	MAE (MW)	MAX AE (MW)	RMSE (MW)	MAPE (%)	MAX APE (%)
Regresja liniowa	29	92	37	3,40	10,36
Sieć MLP	25	67	31	2,93	7,85
Sieć FBF	24	68	31	2,87	8,39

Źródło: opracowanie własne.

Biorąc pod uwagę wszystkie trzy rozważane w tym punkcie zadania prognostyczne, możemy uznać, że w zagadnieniach predykcji krótkoterminowego zapotrzebowania na energię elektryczną sieci neuronowe i systemy rozmyte FBF

są metodami o zbliżonej jakości działania. Tworząc system prognostyczny, należy więc sprawdzić działanie obydwu tych metod, zaś wybór jednej nich uzależnić od konkretnego przypadku i dokonać go na drodze empirycznej. Prognozy otrzymywane za pomocą klasycznych metod prognozowania, opartych na regresji liniowej, dają nieco gorsze wyniki, głównie z powodu nieliniowego charakteru zależności zapotrzebowania na energię od temperatur i innych zmiennych oceniających stan warunków atmosferycznych (patrz uwagi w punkcie 2.1.2).

2.3.3. Systemy z logiką rozmytą typu Takagi–Sugeno

Systemy wnioskowania rozmytego typu Takagi–Sugeno stanowią specyficzną kategorię modeli rozmytych, przeznaczoną w szczególny sposób do zadań identyfikacji systemów. Modele lingwistyczne, opisywane w punkcie 2.3.1, zawierały w konkluzjach (następnikach) reguł (2.3.3) klauzule wykorzystujące zbiory rozmyte, definiujące ograniczenia dla wartości zmiennej wyjściowej. W modelach typu Takagi–Sugeno reguły mają charakter produkcyjny, tzn. ich aktywacja powoduje wykonanie pewnej procedury numerycznej, w postaci wyznaczenia wartości zmiennej wyjściowej określonej funkcji zmiennych wejściowych. Opis lingwistyczny modelu MISO typu Takagi–Sugeno składa się więc z bazy reguł rozmytych następującej postaci:

$$\begin{aligned}&\textbf{JEŻELI } X_1 \text{ jest } A_{11} \textbf{ I } \ldots \textbf{ I } X_n \text{ jest } A_{1n} \quad \textbf{TO } y = f_1(x_1, \ldots, x_n)\\ &\ldots\\ &\textbf{JEŻELI } X_1 \text{ jest } A_{K1} \textbf{ I } \ldots \textbf{ I } X_n \text{ jest } A_{Kn} \quad \textbf{TO } y = f_K(x_1, \ldots, x_n)\end{aligned} \tag{2.3.18}$$

gdzie A_{ij}, j =1,…, n, i = 1, …, K są zbiorami rozmytymi definiującymi wartości zmiennej X_j, dla każdej i-tej reguły, zaś $f_i(\cdot, \ldots, \cdot)$, i = 1, …, K są rzeczywistymi funkcjami zmiennych wejściowych x_1, …, x_n, f_i: $R^n \rightarrow R$, których wartości obliczane są przy aktywacji każdej z reguł systemu.

Proces wnioskowania dla danego wzorca wejściowego x_1, …, x_n przebiega w następujący sposób. Podobnie jak w przypadku zwykłych, lingwistycznych systemów z logika rozmytą, wyznaczana jest wartość stopnia prawdziwości (dopasowania) każdej z reguł τ_i do wzorca x_1, …, x_n:

$$\tau_i = \prod_{j=1}^{n} \mu_{A_{ij}}(x_j), \quad i = 1,\ldots,K \tag{2.3.19}$$

Wyjście systemu wyznaczane jest jako średnia ważona wartości zmiennej wyjściowej obliczonych przez poszczególne reguły:

$$\mathrm{y}(x_1,...,x_n)=\sum_{i=1}^{K} v_i f_i(x_1,...,x_n) \tag{2.3.20}$$

przy czym jako waga wyniku działania każdej reguły v_i przyjmowany jest znormalizowany stopień prawdziwości tej reguły τ_i:

$$v_i=\frac{\tau_i}{\sum_{i=1}^{K}\tau_i}=\frac{\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)}{\sum_{i=1}^{K}\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)},\quad i=1,...,K \tag{2.3.21}$$

Podsumowując więc zależności (2.3.19)–(2.3.21), wyjście systemu rozmytego typu Takagi–Sugeno dla danego wzorca wejściowego x_1, …, x_n wyznaczyć można, korzystając z następującej formuły:

$$\mathrm{y}(x_1,...,x_n)=\frac{\sum_{i=1}^{K} f_i(x_1,...,x_n)\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)}{\sum_{i=1}^{K}\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)} \tag{2.3.22}$$

Zauważmy, że w taki sposób zdefiniowane systemy wnioskowania rozmytego Takagi–Sugeno stanowią rozwinięcie koncepcji uproszczonego wnioskowania rozmytego, zastosowanego w poprzednim punkcie (patrz zależność (2.3.10)) jako podstawa działania modelu neuronowo-rozmytego FBF. Wynik działania systemu FBF zależy wyłącznie od poziomu prawdziwości reguł dla danego wejścia oraz od stałych liczbowych odpowiadających centroidom zbiorów rozmytych następników reguł. W (2.3.22) wykorzystujemy w następnikach reguł ogólniejszą konstrukcję opartą na funkcjach wartości wejściowych.

We wszystkich zastosowaniach opisanych w bieżącym podrozdziale jako funkcje przynależności zbiorów rozmytych w poprzednikach reguł A_{ij}, j =1,…, n, i = 1, …, K, przyjęto, podobnie jak w przypadku modelu FBF, funkcje Gaussa:

$$\mu_{A_{ij}}(x_j)=\exp\left(-\frac{(x_j-a_{ij}^{*})^2}{2\sigma_{ij}^2}\right) \tag{2.3.23}$$

gdzie a_{ij}^{*} i σ_{ij}, j =1,…, n, i = 1, …, K, są podobnie jak w przypadku (2.3.11) parametrami krzywej Gaussa określającej funkcję przynależności zbioru rozmytego A_{ij}.

2.3.4. Prognozowanie zapotrzebowania na energię z wykorzystaniem systemów Takagi–Sugeno z liniowymi następnikami reguł

W bieżącym punkcie zajmiemy się wykorzystaniem do prognozowania zapotrzebowania na energię sieci neuronowo-rozmytej, w której stosuje się wnioskowanie rozmyte typu Takagi–Sugeno, w przypadku gdy funkcje f_i występujące w następnikach reguł są funkcjami liniowymi. Jak zobaczymy, dla tego typu systemów możliwe jest opracowanie efektywnych i stosunkowo prostych metod uczenia wykorzystujących wiedzę o interpretacji parametrów sieci jako odpowiednich elementów systemu rozmytego.

Baza reguł (2.3.18) systemu przyjmie więc następującą postać:

JEŻELI X_1 jest A_{11} **I** . . . **I** X_n jest A_{1n} **TO** $y = m_{10} + m_{11}x_1 + \ldots + m_{1n}x_n$

… (2.3.24)

JEŻELI X_1 jest A_{K1} **I** . . . **I** X_n jest A_{Kn} **TO** $y = m_{K0} + m_{K1}x_1 + \ldots + m_{Kn}x_n$

Przy tym założeniu, dla reguł postaci (2.3.24), na podstawie (2.3.22), równanie systemu możemy zapisać następująco:

$$\mathrm{y}(x_1,\ldots,x_n) = \frac{\sum_{i=1}^{K}(m_{i0} + m_{i1}x_1 + \ldots + m_{in}x_n)\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)}{\sum_{i=1}^{K}\prod_{j=1}^{n}\mu_{A_{ij}}(x_j)} \tag{2.3.25}$$

Zapiszmy (2.3.25) z wykorzystaniem znormalizowanych stopni prawdziwości poszczególnych reguł v_i, $i = 1, \ldots, K$:

$$\mathrm{y}(x_1,\ldots,x_n) = \sum_{i=1}^{K}(m_{i0} + m_{i1}x_1 + \ldots + m_{in}x_n)v_i \tag{2.3.26}$$

gdzie na podstawie (2.3.21) wagę v_i, $i = 1, \ldots, K$, korzystając z założenia o gaussowskim charakterze funkcji przynależności zbiorów rozmytych A_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$, występujących w poprzednikach reguł, możemy wyznaczyć według zależności:

$$v_i = \frac{\prod_{j=1}^{n} \mu_{A_{ij}}(x_j)}{\sum_{i=1}^{K}\prod_{j=1}^{n} \mu_{A_{ij}}(x_j)} = \frac{\prod_{j=1}^{n} \exp\left(\frac{-(x_j - a_{ij}^*)^2}{2\sigma_{ij}^2}\right)}{\sum_{i=1}^{K}\prod_{j=1}^{n} \exp\left(\frac{-(x_j - a_{ij}^*)^2}{2\sigma_{ij}^2}\right)} = \frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^2}\right)}{\sum_{i=1}^{K}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^2}\right)} \quad (2.3.27)$$

Sformułowaliśmy więc działanie systemu rozmytego typu Takagi–Sugeno, z liniowymi następnikami reguł, w postaci modelu (2.3.25) (albo (2.3.26) i (2.3.27)). Do korzystania z systemu niezbędne jest znalezienie jego parametrów: wartości centroidów i szerokości funkcji przynależności zbiorów rozmytych poprzedników każdej reguły: a_{ij}^* i σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$, ($n \cdot K$ parametrów) oraz współczynników funkcji liniowych występujących w następniku każdej reguły: m_{ij}, $j = 0,\ldots, n$, $i = 1, \ldots, K$ (razem $(n + 1)\cdot K$ parametrów, wliczając także wyraz wolny każdej z funkcji).

Do znalezienia wartości parametrów wykorzystamy właściwość, że na analizowany model możemy spojrzeć zarówno jak na neuronową strukturę typu (2.3.25), jak i na system rozmyty, którego parametry mają swoją określoną interpretację semantyczną. Tego typu systemy określane są często jako systemy konekcjonistyczne (czyli oparte na połączeniach), o lokalnej reprezentacji wiedzy. Wiedza w takich modelach ma charakter lokalny, to znaczy wadze każdego z połączeń możemy przypisać określony jej fragment, w przeciwieństwie do np. warstwowej sieci perceptronowej. MLP należą bowiem do systemów o globalnej reprezentacji wiedzy, gdzie wiedza rozkłada się na całościowe wzorce parametrów wagowych sieci i nie daje się jej rozdzielić na poszczególne wagi.

Znalezieniem centroidów i szerokości funkcji przynależności zbiorów rozmytych poprzedników reguł, a_{ij}^* i σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$, zajmiemy się później; wykorzystamy w tym celu różnego rodzaju metody analizy skupień (grupowania danych) w przestrzeni wejść systemu. Obecnie zastanówmy się nad sposobem oszacowania współczynników funkcji liniowych następników każdej z reguł, m_{ij}, $j = 0,\ldots, n$, $i = 1, \ldots, K$, przy założeniu, że a_{ij}^* i σ_{ij} są już znane.

Do oszacowania parametrów następników reguł wykorzystamy metodę najmniejszych kwadratów (MNK). Zauważmy, że wyjście systemu zależy w sposób liniowy od parametrów m_{ij}, więc w celu ich znalezienia nie musimy stosować iteracyjnych metod optymalizacji nieliniowej. Zakładając, że znamy parametry poprzedników reguł a_{ij}^* i σ_{ij}, a co za tym idzie wagi v_i w (2.3.26), możemy zastosować odpowiednią transformację zmiennych i doprowadzić problem do liniowego.

$$\begin{aligned} \mathrm{y}(x_1,...,x_n) &= \sum_{i=1}^{K} v_i m_{i0} + m_{i1} v_i x_1 + ... + m_{in} v_i x_n = \\ &= \sum_{i=1}^{K} z_{i0} m_{i0} + m_{i1} z_{i1} + ... + m_{in} z_{in} \end{aligned} \tag{2.3.28}$$

gdzie:

$$\begin{aligned} z_{ij} &= v_i x_j, \quad j = 1,...,n, i = 1,...,K \\ z_{i0} &= v_i \end{aligned} \tag{2.3.29}$$

stanowi zestaw $(n + 1)\cdot K$ przetransformowanych zmiennych objaśniających nowego, liniowego, problemu MNK.

Przypomnijmy, że współczynniki v_i, $i = 1, \ldots, K$ wyznaczane są z (2.3.27). Do oszacowania parametrów m_{ij} możemy zastosować teraz jedną ze znanych podstawowych metod rozwiązywania zadania regresji liniowej, takich jak metoda równań normalnych albo rozkładu na wartości osobliwe (*Singular Value Decomposition*, SVD).

Uczenie neuronowo-rozmytego systemu Takagi–Sugeno z liniowymi następnikami reguł przebiega więc jako procedura dwuetapowa. Zbiór treningowy składa się ze wzorców wejściowych oraz odpowiadających im znanych (treningowych) wartości zmiennej wyjściowej $\{\mathbf{x}_k, t_k\} = \{(x_{k1}, \ldots, x_{kn}), t_k\}$, $k = 1, \ldots, N$.

1. W pierwszym kroku określane są parametry a_{ij}^* i σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$ funkcji przynależności zbiorów A_{ij} występujących w poprzednikach reguł systemu. Centroidy funkcji Gaussa a_{ij}^* określane są przez zastosowanie algorytmu grupowania danych, w przestrzeni wejść systemu. Wykorzystany tu może zostać w zasadzie dowolny algorytm podziałowego (niehierarchicznego) grupowania danych stosowany w zagadnieniach eksploracji danych, np. algorytm *c*-środków, metoda górska czy algorytm Kohonena. Do uformowania skupień wykorzystywane są wzorce wejściowe ze zbioru treningowego $\mathbf{x}_k = (x_{k1}, \ldots, x_{kn})$, $k = 1, \ldots, N$. Wartości parametrów a_{ij}^* określane są przez centra otrzymanych skupień (grup danych). Liczba tworzonych skupień odpowiada liczbie reguł K i może być ustalana przez algorytm grupowania (przy określeniu liczby reguł modelu) bądź też narzucana z góry (czyli determinowana przez założoną liczbę reguł). Parametry σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$ określane mogą być na podstawie odległości między centrami uformowanych skupień. W wielu przypadkach wyznacza się je jednak w sposób uproszczony, na podstawie zakresu wartości danej zmiennej wejściowej i liczby reguł systemu.

2. W drugim kroku szacowane są parametry współczynników funkcji liniowych następników reguł systemu, m_{ij}, $j = 0,\ldots, n$, $i = 1, \ldots, K$. Proces ten przebiega w zasadzie zgodnie ze standardową procedurą tworzenia modeli

regresji liniowej, z transformacjami zmiennych objaśniających. Dla każdego wzorca wejściowego w danych treningowych $\mathbf{x}_k = (x_{k1}, \ldots, x_{kn})$, $k = 1, \ldots, N$, przy wykorzystaniu (2.3.27), wyznaczane są współczynniki v_i, $i = 1, \ldots, K$. Następnie na podstawie (2.3.29) możemy obliczyć wartości transformowanych zmiennych z_{ij}, $j = 0,\ldots, n$, $i = 1, \ldots, K$, tworząc razem treningową obserwację zmiennej wyjściowej t_k, k-ty wzorzec dla estymacji liniowego modelu MNK: $\{\mathbf{Z}_k, t_k\} = \{z_{kij}, t_k\}, j = 0,\ldots, n, i = 1, \ldots, K, k = 1, \ldots, N$.

Neuronowo-rozmyta sieć Takagi–Sugeno, z liniowymi następnikami reguł, zastosowana została do prognozy szczytowego zapotrzebowania na energię (maksymalnego godzinnego zapotrzebowania na energię) z dwudniowym wyprzedzeniem czasowym (Bartkiewicz 2000d; Bartkiewicz, Butkevych i inni 2001; Butkevych, Pawłowskiy, Bartkiewicz, Zieliński 2002; Bartkiewicz, Bolek i inni 2010). Model zdefiniowany jest przez równanie postaci:

$$\begin{aligned} ZS_d = f(&ZS_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZS(t)_{d-2}, \\ &ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, \ldots, dt_{6d}) \end{aligned} \tag{2.3.30}$$

gdzie oznaczenia są analogiczne jak w modelach prezentowanych poprzednio:

$ZS(t)_d$ – zapotrzebowanie na energię w szczytowej godzinie, w dniu d,

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d, gdzie t oznacza godzinę szczytową,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Jako dane treningowe wykorzystany został dwuletni zbiór obserwacji procesu godzinnego zapotrzebowania na energię oraz pomiarów temperatur w jednej ze spółek dystrybucyjnych. Zbiór testowy obejmował dane z jednego roku. Wykorzystywane dane dotyczyły wszystkich dni tygodnia (włączając soboty i niedziele), ale usunięto z nich obserwacje odnośnie do dni nietypowych (patrz punkt 2.2.5).

W trakcie tworzenia prognozy przeanalizowane zostały dwa modele prognostyczne, wykorzystujące w pierwszym kroku uczenia systemu do oszacowania centroidów zbiorów rozmytych poprzedników reguł różne metody grupowania danych w n-wymiarowej przestrzeni wzorców wejściowych.

TS-Model 1

Do grupowania danych w przestrzeni wejść systemu wykorzystana została tzw. metoda górska. Jest to stosunkowo prosta (podobnie jak zaskakująco wiele innych metod grupowania danych), intuicyjna metoda formowania skupień, niewymagająca wcześniejszego określenia ich liczby. Metoda górska stanowi jedno z tradycyjnych narzędzi wykorzystywanych w identyfikacji systemów

z logiką rozmytą (patrz Yager, Filev 1995; Wu, Lu 1999). Czytelników zainteresowanych dokładniejszą analizą tej metody odsyłamy do literatury przedmiotu. Obecnie przedstawimy jedynie jej najważniejsze elementy.

W pierwszym kroku metody górskiej tworzymy zbiór wszystkich możliwych potencjalnych centrów grupowania C. Jako C możemy przyjąć np. cały zbiór danych treningowych. Innym często stosowanym rozwiązaniem jest wprowadzenie równomiernego podziału każdej zmiennej wejściowej na przedziały (siatki) i przyjęcie jako potencjalne centra grupowania n-wymiarowych wzorców utworzonych z kombinacji krańców (lub środków) tych przedziałów. W naszym przypadku przyjęto pierwsze z tych rozwiązań, więc zbiór C początkowo składa się z N wzorców $\mathbf{x}_k = (x_{k1}, \ldots, x_{kn})$, $k = 1, \ldots, N$.

Następnie dla każdego z potencjalnych centrów grupowania $\mathbf{c}_t \in C$ wyznacza się wartość tzw. funkcji górskiej, określającej potencjał danego wzorca jako centrum skupienia danych.

$$M(\mathbf{c}) = \sum_{k=1}^{N} e^{-\alpha d(\mathbf{x}_k, \mathbf{c})} \tag{2.3.31}$$

gdzie $d(\mathbf{x}_k, \mathbf{c})$ jest miarą odległości między wzorcem $\mathbf{x}_k$ a kandydatem na centrum skupienia $\mathbf{c}$, zazwyczaj (i w naszym przypadku) przyjmuje się miary oparte na metryce euklidesowej, $d(\mathbf{x}_k, \mathbf{c}) = ||\mathbf{x}_k - \mathbf{c}||^2$, zaś α dodatnią stałą.

Łatwo zauważyć, że wartość funkcji górskiej (2.3.31) determinowana jest przez liczbę wzorców danych położonych blisko potencjalnego środka skupiska $\mathbf{c}$, a więc generalnie stanowi ona odzwierciedlenie gęstości rozkładu wzorców danych wokół poszczególnych środków. Im wyższa wartość funkcji górskiej dla danego $\mathbf{c}$, tym więcej danych skupia się w pobliżu $\mathbf{c}$, a więc tym większa jest użyteczność $\mathbf{c}$ jako środka skupienia.

W każdym i-tym kroku, wybierając kolejny środek skupienia, szukamy oczywiście najlepszego kandydata, czyli wzorca o najwyższej wartości funkcji górskiej:

$$\mathbf{c}_{\max} = \arg\max_{\mathbf{c} \in C} |M(\mathbf{c})| \tag{2.3.32}$$

Jeżeli potencjał znalezionego najlepszego kandydata jest zbyt mały (np. $M(\mathbf{c}_{\max}) < \delta$, gdzie δ jest pewną stałą), to wśród kandydatów nie ma już odpowiednich centrów nowych skupisk i należy zakończyć działanie algorytmu. Jeżeli potencjał $\mathbf{c}_{\max}$ jest wystarczający, to dodajemy do systemu nową regułę i przyjmujemy jako centroidy zbiorów rozmytych jej poprzednika współrzędne środka grupowania $\mathbf{c}_{\max}$, czyli $\mathbf{a}_i^* = \mathbf{c}_{\max}$. Następnie przed przejściem do wyboru kolejnego środka skupiska dokonujemy tzw. destrukcji funkcji górskiej.

Destrukcja funkcji górskiej to proces, który ma na celu wyeliminowanie ze zbioru potencjalnych nowych środków skupisk C wpływu znalezionego wcześniej środka $\mathbf{c}_{\max}$. Elementy z C położone blisko znalezionego środka $\mathbf{c}_{\max}$ będą miały również wysoką wartość funkcji górskiej, ponieważ wzorce danych leżące blisko $\mathbf{c}_{\max}$ będą znajdowały się także w ich pobliżu. Aby usunąć ten efekt, dokonuje się korekty zmniejszającej wartości funkcji górskiej poszczególnych elementów C:

$$M(\mathbf{c}) = M(\mathbf{c}) - M(\mathbf{c}_{\max}) \sum_{k=1}^{N} e^{-\beta d(\mathbf{c}_{\max}, \mathbf{c})} \tag{2.3.33}$$

gdzie β jest kolejną stałą dodatnią dobieraną do konkretnego przypadku.

TS-Model 2

W tym modelu do grupowania danych wykorzystano sieć Kohonena (patrz punkt 2.2.6). W przypadku obu badanych modeli parametry szerokości zbiorów rozmytych poprzedników reguł, σ_{ij}, j =1,…, n, i = 1, …, K, dobierane były jako stałe w zależności od zakresu wartości każdej ze zmiennych wejściowych i liczby skupień (reguł systemu) K.

Istotnym elementem w obu przypadkach jest określenie liczby skupień w algorytmie grupowania danych, a co za tym idzie – liczby reguł systemu K. W metodzie górskiej liczba formowanych skupień determinowana jest, co prawda, przez sam algorytm, ale możemy na nią wpływać, dobierając odpowiednio parametry algorytmu. Praktycznie w obu przypadkach liczba skupień (reguł) dobierana musi być na drodze eksperymentalnej poprzez obserwację błędów na walidacyjnym zbiorze danych, tak by uzyskać model o jak najlepszej generalizacji. Interesujące jest, że w przypadku modelu TS-Model 1 (metody górskiej) system najlepszą generalizację uzyskał przy dziesięciu skupieniach (regułach). W przypadku modelu TS-Model 2 (sieci Kohonena) najlepsze wyniki uzyskano przy jedynie czterech regułach (Bartkiewicz 2000d). Liczba skupień ma nie tylko pewne znaczenie dla efektywności działania systemu, ale także dla nakładów niezbędnych do jego tworzenia. Przy 13 zmiennych wejściowych dla 10 reguł należy znaleźć 140 współczynników funkcji liniowych następników, natomiast dla 4 reguł – jedynie 56.

Tabela 2.3.4. Porównanie błędów prognozy szczytowego zapotrzebowania z dwudniowym wyprzedzeniem czasowym dla modeli neuronowo-rozmytych typu Takagi–Sugeno

Model	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
MLP	9 210	42 271	3,73	23,59
TS-Model 1	9 058	46 729	3,61	21,59
TS-Model 2	9 020	47 558	3,63	26,54

Źródło: opracowanie własne.

W tabeli 2.3.4 zaprezentowane zostały wyniki porównujące dokładność prognoz uzyskanych za pomocą obu analizowanych modeli neuronowo-rozmytych typu Takagi–Sugeno oraz sieci neuronowej MLP. Jak widzimy, zarówno w przypadku TS-Model 1, jak i TS-Model 2 otrzymaliśmy podobną dokładność działania systemu prognostycznego, nieznacznie lepszą niż w przypadku sieci MLP. Istotne jest przy tym, że czas potrzebny na identyfikację (uczenie) obu modeli neuronowo-rozmytych był dużo krótszy niż w przypadku treningu sieci MLP.

2.3.5. Prognozowanie zapotrzebowania na energię z wykorzystaniem systemów Takagi–Sugeno z nieliniowymi następnikami reguł

W ogólnym przypadku funkcje $f_i(x_1, \ldots, x_n)$, $i = 1, \ldots, K$, występujące w następnikach reguł w bazie wiedzy systemu rozmytego Takagi–Sugeno (2.3.18), niekoniecznie muszą być funkcjami liniowymi. Mogą być one dowolnymi rzeczywistymi funkcjami zmiennych wejściowych $x_1, \ldots, x_n$.

Do modelowania funkcji następników reguł $f_i(x_1, \ldots, x_n)$ zastosowaliśmy warstwowe sieci perceptronowe MLP (Bartkiewicz 1998a; Bartkiewicz 2011c). Dzięki właściwości uniwersalnej aproksymacji ich użycie nie wymaga określenia *a priori* postaci szacowanej zależności – tak jak w przypadku modeli rozważanych w poprzednim punkcie, gdzie przyjęto jej liniowy kształt.

W procesie budowy modelu wszystkie wzorce treningowe podzielone zostały, przy użyciu algorytmu grupowania danych (sieć Kohonena), na K odrębnych segmentów (gdzie K odpowiada liczbie reguł systemu). Każdy z uzyskanych segmentów posłużył jako zbiór uczący dla odrębnej sieci MLP. Końcowa odpowiedź systemu, dla danego wejścia $x_1, \ldots, x_n$, dana jest przez ogólne równanie modelu wnioskowania typu Takagi–Sugeno (2.3.20):

$$\mathrm{y}(x_1,\ldots,x_n) = \sum_{i=1}^{K} w_i f_i(x_1,\ldots,x_n) \tag{2.3.34}$$

gdzie $f_i(x_1, \ldots, x_n)$, $i = 1, \ldots, K$ oznacza wyjście sieci neuronowej zbudowanej dla i-tego segmentu danych (funkcji następnika i-tej reguły). Współczynniki w_i określane są przez znormalizowane stopnie prawdziwości poszczególnych reguł, czyli stopnie dopasowania danego wejścia $x_1, \ldots, x_n$ do warunków rozmytych zdefiniowanych w poprzedniku danej reguły, przekształcone tak, by spełniały warunek normalizacyjny:

$$\sum_{i=1}^{K} w_i = 1 \tag{2.3.35}$$

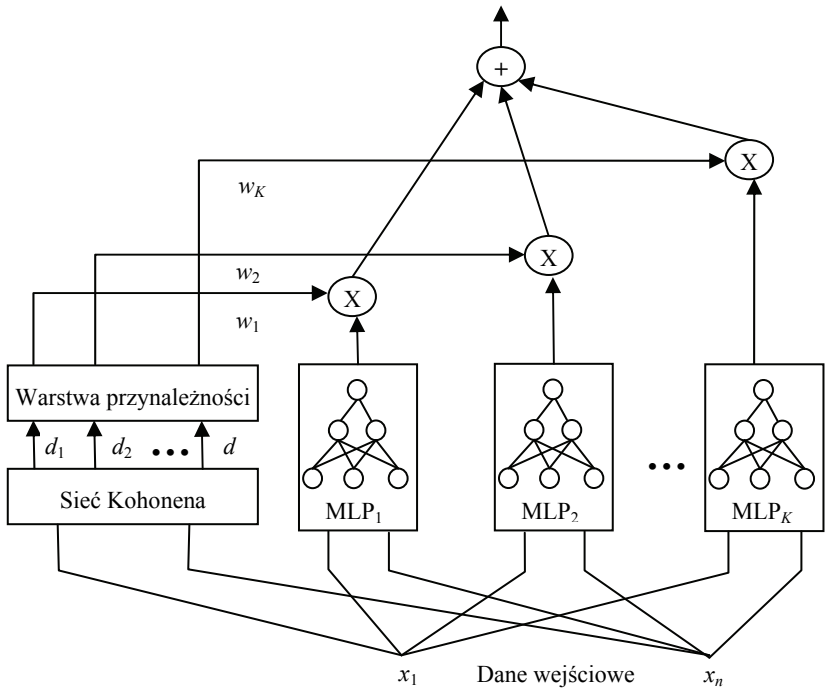
W przypadku badanego modelu zdecydowaliśmy się nie modelować samych warunków rozmytych poprzedników reguł w formie zbiorów rozmytych

A_{ij}, j =1,…, n, i = 1, …, K, lecz wyznaczać wartości współczynników w_i bezpośrednio na podstawie wyników grupowania danych treningowych, wykorzystując wartości funkcji przynależności wzorca wejściowego x_1, …, x_n do i-tego segmentu danych stanowiącego podstawę do budowy sieci neuronowej reguły (rysunek 2.3.2). W tym celu zastosowaliśmy rozmytą wersję sieci Kohonena (Huntsberger, Ajjimarangsee 1990), która obok klasycznej warstwy konkurencyjnej zawiera dodatkową warstwę neuronów wyznaczających stopnie przynależności wektora wejściowego do poszczególnych kategorii.

Przypomnijmy, że w sieci Kohonena (patrz punkt 2.2.6) każdy neuron reprezentuje pojedyncze skupienie (segment) danych, a centroid wyznaczony jest przez jego wektor wagowy. Dla danego wzorca wejściowego $\mathbf{x} = (x_1, \ldots, x_n)$ wartości neuronów sieci określają jego odległości od centroidów (środków) poszczególnych segmentów:

$$d_i = \left\| \mathbf{x} - \mathbf{c}_i \right\| \quad , i = 1,...,K \tag{2.3.36}$$

gdzie K jest liczbą skupień (neuronów), $\mathbf{c}_i$ są wektorami wag neuronów w sieci Kohonena (centroidami poszczególnych grup).



Rysunek 2.3.2. Sieć neuronowo-rozmyta typu Takagi–Sugeno z nieliniowymi funkcjami w następnikach reguł w postaci warstwowych sieci neuronowych
Źródło: opracowanie własne

W klasycznym modelu sieci Kohonena o reakcji całej sieci (wyborze segmentu, do którego kwalifikowany jest wzorzec wejściowy) decyduje tzw. neuron zwycięski, to znaczy taki, którego wartość wyjściowa (odległość d_i) jest najmniejsza. Huntersberger i Ajjimarangsee w rozmytej wersji wprowadzają dodatkową warstwę neuronów obliczającą stopnie przynależności wektora wejściowego do wszystkich segmentów reprezentowanych przez neurony sieci. Wykorzystali oni w tym celu procedurę zbudowaną przez Bezdeka na potrzeby rozmytych algorytmów grupowania ISODATA (c-środków).

Wartość τ_i określająca siłę przynależności wzorca wejściowego $\mathbf{x} = (x_1, \ldots, x_n)$ do i-tej grupy danych (siłę dopasowania $\mathbf{x}$ do warunków i-tej reguły) wyznaczana jest na podstawie stanów wyjściowych neuronów w warstwie Kohonena, d_i (odległości $\mathbf{x}$ od centroidów poszczególnych segmentów):

$$\tau_i = \left(\sum_{k=1}^{K} \left(\frac{d_i}{d_k} \right)^p \right)^{-1}, i = 1,..., K \tag{2.3.37}$$

Idea (2.3.37) jest dosyć oczywista. Jeżeli dla danego wzorca wejściowego $\mathbf{x}$ i-ty neuron w warstwie Kohonena „wygrywa" wyraźnie z innymi, oznacza to, że $d_i << d_k$, dla wszystkich k. Wówczas stosunki d_i/d_k są liczbami małymi, bliskimi 0. Siła przynależności τ_i, jako odwrotność ich sumy, musi więc być dosyć wysoka. Jeżeli rozkład przynależności $\mathbf{x}$ do poszczególnych segmentów jest bardziej równomierny, czyli neuron zwycięski „wygrywa" w porównaniu z innymi jedynie nieznacznie, to dla pewnych k stosunki $d_i/d_k \approx 1$. Odwrotność ich sumy musi więc być liczbą stosunkowo małą, zbliżającą się do 0.

Parametr p określa, jak szybko siła przynależności wzorca wejściowego $\mathbf{x}$ do danego segmentu ma maleć w miarę wzrostu odległości tego wzorca od środka segmentu. Dobierany jest on do konkretnego przypadku.

Współczynniki w_i we wzorze (2.3.34) przyjmowane są jako znormalizowane stopnie przynależności wzorca wejściowego $\mathbf{x} = (x_1, \ldots, x_n)$ do poszczególnych segmentów danych:

$$w_i = \frac{\tau_i}{\sum_{i=1}^{K} \tau_i}, \quad i = 1,..., K \tag{2.3.38}$$

Zaprezentowany model, stanowiący połączenie systemów rozmytych typu Takagi–Sugeno i warstwowych sieci perceptronowych MLP (TS–MLP), zastosowany został do omawianej w poprzednim punkcie 2.3.4 prognozy szczytowego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym, określonej przez (2.3.30) (Bartkiewicz 1998a). Do jego tworzenia

i testowania wykorzystane zostały te same zbiory danych, co w przypadku modeli neuronowo-rozmytych typu Takagi–Sugeno, z liniowymi funkcjami następników reguł.

Model TS–MLP obejmował cztery reguły (podobnie jak model TS-Model 2), więc jego stworzenie wymagało treningu czterech sieci MLP (Bartkiewicz 2011c). Błędy prognoz otrzymanych dla modelu TS–MLP znajdują się w tabeli 2.3.5. Zamieszczono w niej również dla porównania wyniki testowania modeli z liniowymi następnikami reguł oraz pojedynczej sieci MLP, prezentowane już wcześniej w tabeli 2.3.4. Jak widzimy, model TS–MLP osiągnął wyraźnie lepszą dokładność niż pojedyncza sieć neuronowa i nieco lepszą niż sieci neuronowo-rozmyte z liniowymi funkcjami następników. Należy jednak zwrócić uwagę, że uczenie systemów prognostycznych tego typu wymaga bardzo dużych nakładów obliczeniowych, zwłaszcza w porównaniu z TS-Model 1 i TS-Model 2.

Tabela 2.3.5. Porównanie błędów prognozy szczytowego zapotrzebowania z dwudniowym wyprzedzeniem czasowym dla modeli neuronowo-rozmytych typu Takagi–Sugeno z liniowymi i nieliniowymi następnikami reguł

Model	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
MLP	9 210	42 271	3,73	23,59
TS-Model 1	9 058	46 729	3,61	21,59
TS-Model 2	9 020	47 558	3,63	26,54
TS–MLP	8 805	43 410	3,48	20,05

Źródło: opracowanie własne.

Zauważmy jeszcze pewne powiązanie modelu TS–MLP z omawianym w punkcie 2.2.7 podejściem opartym na grupie lokalnych sieci MLP, przełączanych przy użyciu klasyfikatora. Przypomnijmy, że metoda ta również polegała na pogrupowaniu zbioru wzorców treningowych oraz na stworzeniu odrębnego modelu sieci MLP dla każdego z otrzymanych segmentów. Odmiennie jednak niż w przypadku TS–MLP, dla danego wzorca wejściowego **x**, dokonuje się klasyfikacji tego wzorca do odpowiedniego segmentu danych, a co za tym idzie – wyboru jednej sieci MLP wykorzystywanej do uzyskania prognozy. W badaniach w punkcie 2.2.7 do grupowania danych i klasyfikacji nowych wzorców wykorzystaliśmy klasyczną (nierozmytą) sieć Kohonena.

Tradycyjne techniki grupowania i klasyfikacji wykorzystujące metodę „zwycięzca bierze wszystko” (tzn. w przypadku naszego problemu dla konkretnej prognozy wybieramy jeden konkretny model MLP odpowiadający segmentowi, do którego zakwalifikowany został wzorzec wejściowy) nie uwzględniają

sytuacji, w której dane wejściowe leżą blisko granicy decyzyjnej między poszczególnymi segmentami, w dużej odległości od środka (centroidu) kategorii, do której zostały zaliczone. Wykorzystanie klasyfikatora rozmytego zwiększa odporność modelu na błędnie zakwalifikowane wzorce.

W przypadku klasyfikatora ostrego (nierozmytego) błędna klasyfikacja wzorca wejściowego powoduje wybór niewłaściwego modelu sieci perceptronowej oraz, w konsekwencji, prognozę odległą od rzeczywistości. W przypadku klasyfikacji rozmytej, takiej jak w TS–MLP, degradacja działania modelu ma charakter stopniowy, zwłaszcza że błędy dotyczą przede wszystkim wzorców leżących w pobliżu granicy decyzyjnej między sąsiednimi segmentami. Umożliwia ona również naturalne uwzględnienie faktu słabego dopasowania tego typu danych do jednego konkretnego skupienia. Jeśli wzorzec należy w pewnym stopniu do większej liczby segmentów, każda z odpowiadających im sieci perceptronowych będzie miała udział w wyniku końcowym modelu (Bartkiewicz 1998a).

Do weryfikacji tej hipotezy wykonano badania porównawcze neuronowo-rozmytego modelu TS–MLP z prezentowanym w punkcie 2.2.7 podejściem opartym na zespole lokalnych sieci MLP i ostrej (nierozmytej) klasyfikacji wzorca wejściowego. Badania wykonano dla takiej samej jak poprzednio prognozy szczytowego zapotrzebowania z dwudniowym wyprzedzeniem czasowym (2.3.30), lecz w przypadku innego systemu elektroenergetycznego. Otrzymane wyniki przedstawione zostały w tabeli 2.3.6.

Tabela 2.3.6. Porównanie błędów prognozy szczytowego zapotrzebowania z dwudniowym wyprzedzeniem czasowym dla modelu z ostrą i rozmytą klasyfikacją wzorca wejściowego

Model	MAE (kWh)	MAX AE (kWh)	RMSE (kWh)	MAPE (%)	MAX APE (%)
Sieć MLP	10 715	55 755	14 288	4,65	31,26
Zespół lokalnych MLP	10 236	55 340	14 131	4,35	31,79
TS–MLP	9 862	53 678	13 446	4,19	31,66

Źródło: opracowanie własne.

Jak widzimy, badania praktyczne potwierdzają wcześniej postawioną hipotezę. Dokładność działania modelu opartego na rozmytej klasyfikacji i wyborze sieci jest wyraźnie wyższa niż w przypadku ostrego wyboru jednego z lokalnych modeli. Również i w tym przypadku zastosowanie modelu TS–MLP dało lepsze wyniki prognozy niż pojedyncza sieć neuronowa.

2.4. Podsumowanie

W bieżącym rozdziale przeanalizowaliśmy pokrótce problematykę krótkoterminowego prognozowania zapotrzebowania na energię elektryczną. Naszym celem było wskazanie potrzeby stosowania w tej dziedzinie indukcyjnych metod modelowania nieliniowego, do których zaliczyć możemy sieci neuronowe i neuronowo-rozmyte, a także przedstawienie podstawowych zagadnień związanych z ich tworzeniem oraz doświadczeń autora w tej dziedzinie.

Jak pokazaliśmy, w przypadku krótkoterminowej prognozy zapotrzebowania na energię podstawowymi czynnikami wpływającymi na jego wielkość są wartości historyczne obciążeń z przeszłości (z okresu kilku ostatnich dni poprzedzających prognozę) oraz informacje na temat warunków pogodowych w czasie predykcji. Przede wszystkim ostatnia grupa czynników wykazuje nieliniową charakterystykę wpływu na wielkość zapotrzebowania, nie dostarczając jednocześnie żadnych obserwowalnych wzorców kształtu tej nieliniowości (punkt 2.1.2). W związku z tym, jak pokazaliśmy, użycie metod indukcyjnych jest niezbędne zwłaszcza w przypadku modeli korzystających z wejściowej informacji meteorologicznej.

W bieżącym rozdziale przyjrzeliśmy się wykorzystaniu do modelowania krótkookresowego zapotrzebowania na energię elektryczną podstawowych rodzajów sieci neuronowych i neuronowo-rozmytych, powszechnie stosowanych w zadaniach prognostycznych. Wśród modeli neuronowych skoncentrowaliśmy się przede wszystkim na warstwowych sieciach perceptronowych (MLP) (podrozdział 2.2). Innym rodzajem sieci neuronowej wykorzystywanym w zagadnieniach krótkoterminowej prognozy zapotrzebowania na energię jest sieć z funkcjami o bazie radialnej (RBF). Modele tego rodzaju są jednak funkcjonalnie równoważne rozważanym przez nas sieciom neuronowo-rozmytym FBF (punkt 2.3.2). W zakresie modeli neuronowo-rozmytych skupiliśmy się na najczęściej wykorzystywanych w praktycznych zagadnieniach prognozowania addytywnych modelach rozmytych z regułą wnioskowania Larsena. Przyjrzeliśmy się przy tym całej rodzinie sieci z tej grupy, począwszy od modeli FBF ze stałymi wartościami następników reguł (punkt 2.3.2), poprzez modele typu Takagi–Sugeno z liniowymi funkcjami następników (punkt 2.3.4), aż do modeli z nieliniowymi funkcjami następników, w formie sieci perceptronowych MLP (punkt 2.3.5).

Analizując w bieżącym rozdziale szereg zadań prognostycznych z zakresu modelowania krótkoterminowego zapotrzebowania na energię, pokazaliśmy lepsze działanie sieci neuronowych i neuronowo-rozmytych w stosunku do klasycznych metod prognostycznych opartych na liniowych modelach statystycznych (modelach regresji liniowej), co pozwala uzasadnić elementy tezy naszej pracy związane z przydatnością tego rodzaju rozwiązań.

Porównując działanie różnych sieci neuronowych i neuronowo-rozmytych, należy podkreślić, że stosunkowo najlepsze wyniki osiągnęliśmy w przypadku sieci typu Takagi–Sugeno z nieliniowymi następnikami reguł. Jednakże poprawa dokładności prognozy zapotrzebowania jest na tyle nieznaczna, że nie do końca uzasadnia stosowanie tych skomplikowanych i kosztownych obliczeniowo modeli. Godnym rekomendacji podejściem wydają się, z kolei, sieci neuronowo-rozmyte typu Takagi–Sugeno z liniowymi funkcjami następników. W badanych zadaniach wykazywały one nieco lepszą dokładność prognozy zapotrzebowania na energię niż sieci MLP czy FBF. Różnice są jednak na tyle nieduże, że te trzy podejścia należy uznać za równoważne metody prognostyczne, a ich wybór – uzależnić od konkretnego zadania.

Należy naturalnie przypomnieć, że rozdział ten ma charakter pomocniczy. Prezentowana rozprawa poświęcona jest analizie niepewności neuronowych i neuronowo-rozmytych prognoz krótkoterminowego zapotrzebowania na energię elektryczną i ocenie ryzyka związanego z tą niepewnością dla problemów decyzyjnych występujących na rynkach energii. Z tego powodu charakterystyka samej technologii prognostycznej jako takiej ograniczona została do minimum.

ROZDZIAŁ 3

Modelowanie niepewności neuronowych i neuronowo-rozmytych prognoz zapotrzebowania na energię

Prace związane z zastosowaniem nieliniowych metod prognozowania, neuronowych i neuronowo-rozmytych (podobnie zresztą jak również w przypadku innych narzędzi analizy statystycznej), do krótkoterminowej prognozy zapotrzebowania na energię koncentrują się głównie na prognozach punktowych, tj. uzyskaniu konkretnej wartości przewidywanej wielkości. Użytkownicy mają tendencję do traktowania wykorzystywanego modelu prognostycznego jako cudownej czarnej skrzynki, z której otrzymują nieomylne i pewne informacje dotyczące przyszłych faktów.

Tymczasem takie podejście nie uwzględnia jednego istotnego czynnika. **Prognoza nie stanowi informacji nieomylnej i pewnej**. Rzeczywista realizacja procesu zapotrzebowania na poziomie dokładnie zgodnym z prognozą jest z oczywistych względów niemal niemożliwa. Każda prognoza musi być obarczona błędem, w związku z tym wykorzystując ją w problemach decyzyjnych, należy być świadomym niepewności, jaka się z nią wiąże, oraz wynikającym z tej niepewności ryzykiem podejmowanych decyzji.

Można naturalnie powiedzieć, że kwestia dokładności uwzględniana jest w postaci różnego rodzaju błędów prognozy. Ocena jakości działania modelu na podstawie statystycznych miar dopasowania do prób uczących lub danych testowych, opartych na miarach błędu kwadratowego, bezwzględnego lub procentowego, które nie są bezpośrednio powiązane z procesami decyzyjnymi zasilanymi przez system prognostyczny, powoduje jednak, że często szacowane są niewłaściwe aspekty problemu prognostycznego. Łatwo można pokazać, że wielkość zamówienia energii na rynku, optymalna z punktu widzenia minimalizacji oczekiwanego błędu kwadratowego prognozy, w pewnych warunkach niekoniecznie musi być wielkością najlepszą z punktu widzenia ryzyka finansowego podmiotu uczestniczącego w rynku (czym m.in. zajmiemy się w rozdziale 4).

W praktyce, z punktu widzenia procesu decyzyjnego, znajomość wyłącznie prognozowanej, punktowej wartości zapotrzebowania na energię często może

być niewystarczająca. Zgodnie z postawioną w naszej pracy tezą, do oszacowania ryzyka podejmowanych decyzji niezbędne jest przeanalizowanie niepewności prognozy, do czego potrzebne są dodatkowe informacje otrzymywane z całego wynikowego rozkładu warunkowego prognozowanego procesu (przedziały wiarygodności prognozy, gęstość lub dystrybuanta jej rozkładu itd.).

Dlatego, by wykazać prawdziwość tej tezy, obecnie zajmiemy się dokładniejszą analizą wyjścia neuronowego (neuronowo-rozmytego) systemu prognozy zapotrzebowania na energię, czynnikami wpływającymi na jego niepewność oraz metodami określania rozkładu prawdopodobieństwa zapotrzebowania dla danej wartości wejścia modelu. Przedstawimy przy tym wyniki badań eksperymentalnych dotyczących weryfikacji metod szacowania niepewności (wariancji lub odchylenia standardowego) wyjścia, większości prezentowanych w poprzednim rozdziale modeli neuronowych i neuronowo-rozmytych, dla różnych zadań z zakresu krótkoterminowego prognozowania zapotrzebowania na energię elektryczną. Potwierdzają one postawioną w pracy hipotezę badawczą o właściwym działaniu tych narzędzi dla analizowanych architektur neuronowych i neuronowo-rozmytych oraz o możliwości ich zastosowania w praktycznych systemach prognostycznych z badanej dziedziny.

W bieżącym rozdziale interesować nas będzie głównie probabilistyczna ocena niepewności modelu, metody szacowania prawdopodobieństwa osiągnięcia określonych poziomów zapotrzebowania na energię oraz przedziałów jego prognozy (przedziałów wartości, w których zapotrzebowanie znajdzie się z określonym prawdopodobieństwem). W następnym rozdziale przyjrzymy się z kolei, w jaki sposób niepewność prognozy przenosi się na decyzje związane z obrotem energią, przede wszystkim na planowanie transakcji na rynku i strategie postępowania dystrybutorów energii.

3.1. Błąd kwadratowy i interpretacja modelu prognostycznego

3.1.1. Wyjście nieliniowego modelu prognostycznego

W pierwszym kroku spróbujmy skoncentrować się na samym wyjściu modelu prognostycznego. Zastanówmy się nad charakterem prognozy otrzymywanej z systemu – wartości, którą otrzymujemy z systemu i na której będziemy opierać nasze decyzje.

Określmy najpierw pewne podstawowe warunki, w jakich wykonujemy naszą analizę. Przyjmijmy, że korzystamy z nieliniowego modelu w postaci sieci neuronowej lub neuronowo-rozmytej $f(\mathbf{x}, \mathbf{w})$ o charakterze regresyjnym, tzn. modelującej pewne odwzorowanie stochastyczne między ciągłą zmienną

objaśnianą y (zapotrzebowaniem na energię elektryczną lub moc) a wielowymiarową zmienną objaśniającą $\mathbf{x} = (x_1, \ldots, x_n)$ (zmiennymi wejściowymi modelu)

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon \tag{3.1.1}$$

gdzie $\mathbf{w}$ jest zbiorem parametrów (wag) modelu, zaś ε addytywnym czynnikiem losowym. Załóżmy dalej, że oszacowanie parametrów modelu nastąpiło na zbiorze danych treningowych D składającym się ze wzorców wejściowych oraz odpowiadających im znanych (treningowych) wartości zmiennej wyjściowej:

$$D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}, k = 1, \ldots, N \tag{3.1.2}$$

W zasadzie we wszystkich rozważanych w poprzednim rozdziale modelach ostateczna prognoza określana jest przez model trenowany metodą najmniejszych kwadratów, tzn. poprzez minimalizację błędu kwadratowego na zbiorze treningowym D:

$$Err = \frac{1}{2N}\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}))^2 \tag{3.1.3}$$

Szukamy modelowanej zależności w pewnej populacji ogólnej, zakładając, że zbiór treningowy jest dostatecznie dobrą jej reprezentacją. Cel uczenia polega więc na określeniu modelu minimalizującego błąd dla nieskończonego zbioru obserwacji. Jeśli rozważymy błąd kwadratowy modelu uzyskanego na zbiorze treningowym o długości N, przy N dążącym do nieskończoności, to stanowi on oszacowanie wartości oczekiwanej kwadratu odchylenia pomiędzy wartościami zmiennej y i wyjścia sieci $f(\mathbf{x}, \mathbf{w})$ we wspólnym rozkładzie prawdopodobieństwa zmiennych y i $\mathbf{x}$:

$$Err = \lim_{N\to\infty}\frac{1}{2N}\sum_{i=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}))^2 = \frac{1}{2}\iint_R (y - f(\mathbf{x}, \mathbf{w}))^2 p(y, \mathbf{x}) dy d\mathbf{x} \tag{3.1.4}$$

gdzie $p(y, \mathbf{x})$ jest gęstością wspólnego rozkładu prawdopodobieństwa zmiennych y i $\mathbf{x}$.

Korzystając z definicji prawdopodobieństwa warunkowego: $p(y / \mathbf{x}) = p(y, \mathbf{x}) / p(\mathbf{x})$, możemy błąd (3.1.4) zapisać jako:

$$Err = \frac{1}{2}\iint_R (y - f(\mathbf{x}, \mathbf{w}))^2 p(y / \mathbf{x}) p(\mathbf{x}) dy d\mathbf{x} \tag{3.1.5}$$

Oznaczmy przez $E(y / \mathbf{x})$ oraz $E(y^2 / \mathbf{x})$ wartości oczekiwane odpowiednio zmiennej y oraz y^2 w rozkładzie warunkowym $p(y / \mathbf{x})$:

$$E(y/\mathbf{x}) = \int_{-\infty}^{\infty} y p(y/\mathbf{x}) dy \tag{3.1.6a}$$

$$E(y^2/\mathbf{x}) = \int_{-\infty}^{\infty} y^2 p(y/\mathbf{x}) dy \tag{3.1.6b}$$

W zależności (3.1.5) dodając i odejmując $E(y / \mathbf{x})$, a następnie korzystając ze wzoru na kwadrat sumy i addytywności całki, otrzymujemy:

$$\begin{aligned} Err &= \frac{1}{2}\iint_R (y - E(y/\mathbf{x}) + E(y/\mathbf{x}) - f(\mathbf{x},\mathbf{w}))^2 p(y/\mathbf{x})\, p(\mathbf{x}) dy d\mathbf{x} = \\ &= \frac{1}{2}\iint_R (y - E(y/\mathbf{x}))^2 p(y/\mathbf{x})\, p(\mathbf{x}) dy d\mathbf{x} + \\ &+ \iint_R (y - E(y/\mathbf{x}))(E(y/\mathbf{x}) - f(\mathbf{x},\mathbf{w})) p(y/\mathbf{x})\, p(\mathbf{x}) dy d\mathbf{x} + \\ &+ \frac{1}{2}\iint_R (E(y/\mathbf{x}) - f(\mathbf{x},\mathbf{w}))^2 p(y/\mathbf{x})\, p(\mathbf{x}) dy d\mathbf{x} \end{aligned} \tag{3.1.7}$$

Zauważmy, że $E(y / \mathbf{x})$ jako wartość oczekiwana zmiennej y, przy danym $\mathbf{x}$, jak i wyjście modelu $f(\mathbf{x}, \mathbf{w})$ są wielkościami stałymi względem zmiennej y (zależą tylko od $\mathbf{x}$). Porządkując więc (3.1.7) i wyciągając przed całkę względem y elementy niezależne od tej zmiennej, otrzymujemy:

$$\begin{aligned} Err &= \frac{1}{2}\int_{-\infty}^{\infty} (E(y/\mathbf{x}) - f(\mathbf{x},\mathbf{w}))^2 \left(\int_{-\infty}^{\infty} p(y/\mathbf{x})\, dy \right) p(\mathbf{x}) d\mathbf{x} + \\ &+ \int_{-\infty}^{\infty} (E(y/\mathbf{x}) - f(\mathbf{x},\mathbf{w})) \left(\int_{-\infty}^{\infty} y p(y/\mathbf{x})\, dy - E(y/\mathbf{x}) \int_{-\infty}^{\infty} p(y/\mathbf{x})\, dy \right) p(\mathbf{x}) d\mathbf{x} + \\ &+ \frac{1}{2}\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} (y - E(y/\mathbf{x}))^2 p(y/\mathbf{x})\, dy \right) p(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{3.1.8}$$

Zauważmy, że w (3.1.8) odwróciliśmy również kolejność poszczególnych głównych członów składających się na (3.1.7). Przyjrzyjmy się teraz bliżej środkowemu członowi zależności (3.1.8). Biorąc pod uwagę (3.1.6a) oraz fakt, że z definicji funkcji gęstości prawdopodobieństwa $\int_{-\infty}^{\infty} p(y/\mathbf{x})\, dy = 1$, otrzymujemy:

$$\int_{-\infty}^{\infty}(E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))\left(\int_{-\infty}^{\infty}yp(y/\mathbf{x})\,dy-E(y/\mathbf{x})\int_{-\infty}^{\infty}p(y/\mathbf{x})\,dy\right)p(\mathbf{x})d\mathbf{x}=$$
$$=\int_{-\infty}^{\infty}(E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))\big(E(y/\mathbf{x})-E(y/\mathbf{x})\cdot 1\big)p(\mathbf{x})d\mathbf{x}=0 \tag{3.1.9}$$

Środkowy człon zależności (3.1.8) jest więc równy 0. Pierwszy człon (3.1.8) możemy uprościć jedynie nieznacznie, ponownie korzystając z właściwości, że całka funkcji gęstości rozkładu prawdopodobieństwa w całej przestrzeni jest równa 1:

$$\frac{1}{2}\int_{-\infty}^{\infty}(E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))^2\left(\int_{-\infty}^{\infty}p(y/\mathbf{x})\,dy\right)p(\mathbf{x})d\mathbf{x}=$$
$$=\frac{1}{2}\int_{-\infty}^{\infty}(E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))^2\,p(\mathbf{x})d\mathbf{x} \tag{3.1.10}$$

Zauważmy dalej, że w trzecim członie (3.1.8) wewnętrzna całka (ujęta w nawiasy), biorąc pod uwagę definicję wartości oczekiwanej (3.1.6b), równa jest wartości oczekiwanej zmiennej $(y - E(y/\mathbf{x}))^2$, w warunkowym rozkładzie prawdopodobieństwa $p(y/\mathbf{x})$. Człon ten więc możemy zapisać następująco:

$$\frac{1}{2}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}(y-E(y/\mathbf{x}))^2\,p(y/\mathbf{x})\,dy\right)p(\mathbf{x})d\mathbf{x}=$$
$$=\frac{1}{2}\int_{-\infty}^{\infty}E((y-E(y/\mathbf{x}))^2/\mathbf{x})\,p(\mathbf{x})d\mathbf{x} \tag{3.1.11}$$

Podsumowując zatem, na podstawie wzorów (3.1.9)–(3.1.11) możemy granicę błędu kwadratowego *Err* (3.1.8), przy liczbie wzorców treningowych dążącej do nieskończoności, zapisać jako:

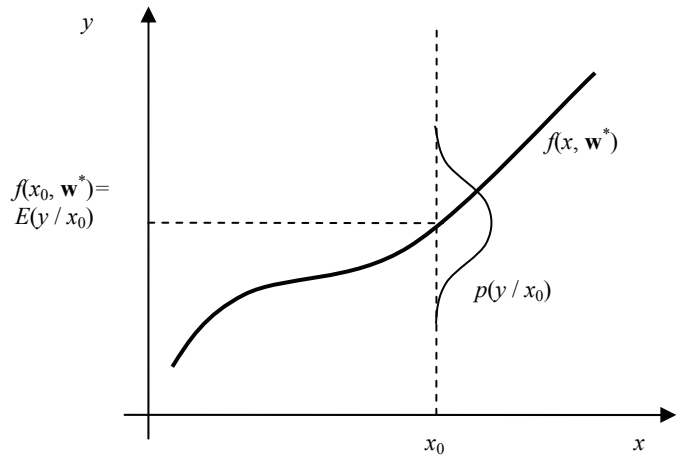
$$Err=\frac{1}{2}\int_{-\infty}^{\infty}(E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))^2\,p(\mathbf{x})d\mathbf{x}+$$
$$+\frac{1}{2}\int_{-\infty}^{\infty}E((y-E(y/\mathbf{x}))^2/\mathbf{x})\,p(\mathbf{x})d\mathbf{x} \tag{3.1.12}$$

Celem uczenia modelu neuronowego lub neuronowo-rozmytego metodą najmniejszych kwadratów jest znalezienie zestawu wag (parametrów) modelu $\mathbf{w}^*$, minimalizującego błąd kwadratowy (3.1.12). Zauważmy przy tym, że drugi człon (3.1.12) nie zależy od wag modelu **w**. Z punktu widzenia minimalizacji ma on więc charakter stały i można go zaniedbać. Łatwo więc dostrzec, że asympto-

tyczny błąd (3.1.12) przyjmuje globalną wartość minimalną wtedy, gdy pierwszy człon zanika, tj. dla optymalnego zestawu parametrów $\mathbf{w}^*$ mamy:

$$f(\mathbf{x}, \mathbf{w}^*) = E(y / \mathbf{x}) \tag{3.1.13}$$

Wykonana więc tutaj standardowa analiza błędu kwadratowego i uczenia metodą najmniejszych kwadratów doprowadziła nas do ważnego wniosku odnośnie do interpretacji prognozy (wyjścia modelu prognostycznego). Dla optymalnego zestawu parametrów, w punkcie minimum błędu kwadratowego wyjście modelu $f(\mathbf{x}, \mathbf{w}^*)$ równe jest wartości oczekiwanej prognozowanej zmiennej y, dla danego wzorca wejściowego $\mathbf{x}$. W statystyce funkcję $y(\mathbf{x}) = E(y / \mathbf{x})$ określa się często **funkcją regresji** zmiennej y względem zmiennej wejściowej $\mathbf{x}$. **Dla określonego zestawu danych wejściowych prognoza zapotrzebowania na energię (lub moc) elektryczną otrzymywana z modelu stanowi oszacowanie wartości oczekiwanej (regresji) tego zapotrzebowania, dla wykorzystanych danych wejściowych**.



Rysunek 3.1.1. Schematyczna ilustracja właściwości (3.1.13)
Źródło: opracowanie własne

Schematyczna ilustracja tego faktu znajduje się na rysunku 3.1.1. Prezentujemy na nim przykładowe odwzorowanie $R \rightarrow R$, realizowane przez sieć neuronową lub neuronowo-rozmytą minimalizującą błąd kwadratowy, określające warunkową wartość oczekiwaną rozkładu zmiennej wyjściowej y. Dla dowolnej wartości zmiennej wejściowej x_0, wyjście sieci $f(x_0, \mathbf{w}^*)$ dane jest przez funkcję regresji, czyli wartość oczekiwaną (średnią) zmiennej objaśnianej (wyjściowej) y, w odniesieniu do rozkładu prawdopodobieństwa $p(y / x_0)$ tej zmiennej objaśnianej, dla powyższej wartości x_0.

To ważna informacja zarówno z punktu widzenia zwykłego wykorzystania modelu, jak i jego dalszej analizy pod kątem wpływu prognozy na dalsze procesy decyzyjne. Z jednej strony mówi nam o tym, że otrzymywana prognoza jest najbardziej prawdopodobną, najpewniejszą wartością zapotrzebowania dla posiadanego zestawu informacji wejściowych. W związku z tym w dalszych działaniach najbezpieczniej, najpewniej kierować się tą właśnie wartością. Z drugiej jednak strony, prognoza określana przez model to tylko wartość oczekiwana rozkładu prawdopodobieństwa dla danych wejściowych. Stanowi ona istotną, ale tylko wybraną charakterystykę tego rozkładu. Wskazuje to, że istnieją jeszcze inne, dodatkowe informacje na temat prognozowanego zapotrzebowania, które zaniedbujemy. Być może, dobrym pomysłem jest przyjrzeć się im i spróbować je wykorzystać.

Jakość zaprezentowanego oszacowania funkcji regresji zależy oczywiście od szeregu warunków, które należy spełnić w trakcie uczenia sieci neuronowej lub neuronowo-rozmytej. Po pierwsze, zbiór uczący D musi być dostatecznie duży (albo reprezentatywny), aby błąd kwadratowy dla skończonego zestawu danych treningowych pozwalał na dobre przybliżenie błędu granicznego. Po drugie, model musi być w stanie osiągnąć minimum błędu. Musi istnieć taki zestaw wag (parametrów) sieci neuronowej lub neuronowo-rozmytej, dla którego model będzie w stanie osiągnąć wartości bliskie optimum. Po trzecie, metoda uczenia musi zapewniać osiągnięcie tego minimum.

Zasygnalizowane trzy podstawowe problemy, związane z jakością działania modelu, będziemy dokładniej analizować w kolejnym punkcie. Przekonamy się, że sytuacja nie jest taka oczywista; źródła błędu modelu mają często charakter nawzajem konfliktowy, sprzeczny, istnieje pomiędzy nimi pewna wymienność. Ponadto otwarty, jak na razie, pozostaje problem wspomniany wcześniej – co czasami powinniśmy wiedzieć o prognozie zapotrzebowania obok jego wartości oczekiwanej.

3.1.2. Źródła niepewności modeli neuronowych i neuronowo-rozmytych

W poprzednim punkcie, analizując proces uczenia nieliniowego modelu prognostycznego, pokazaliśmy, że odwzorowanie realizowane przez optymalną sieć neuronową lub neuronowo-rozmytą $y = f(\mathbf{x}, \mathbf{w})$, minimalizującą błąd kwadratowy na zbiorze treningowym, w miarę wzrostu liczby wzorców uczących N dąży do $E(y / \mathbf{x})$, czyli do wartości oczekiwanej zmiennej objaśnianej y, dla danego wzorca wejściowego $\mathbf{x}$.

Jeżeli jednak przyjrzymy się zależności (3.1.12), która doprowadziła nas do tego wniosku, łatwo zauważymy, że nawet idealna aproksymacja wartości oczekiwanej przez model prognostyczny nie zapewnia redukcji błędu kwadratowego do zera. W idealnym przypadku zeruje się pierwszy komponent (3.1.12),

pozostaje natomiast drugi, który określiliśmy jako stały z punktu widzenia optymalizacji. Jako że stały, jest on nieredukowalny. Komponent ten, wyznaczający dolną granicę błędu w procesie optymalizacji wag sieci, reprezentuje element losowy, nieodłączny szum w danych. Ponadto pamiętać należy o kwestii generalizacji, delikatnej różnicy między modelem granicznym a faktycznym.

Obecnie więc postawimy sobie zbliżone, ale jednak nieco odmienne pytanie niż w poprzednim punkcie. Przyjmijmy, że mamy model $y = f(\mathbf{x}, \mathbf{w})$, minimalizujący błąd kwadratowy na zbiorze treningowym $D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$. Powstaje pytanie, jaka jest dla konkretnej prognozy (konkretnego wejścia modelu) spodziewana (średnia) wartość odchylenia między wyjściem modelu a faktyczną wartością prognozowanego zapotrzebowania. Dokładniej, aby pozbyć się znaku odchylenia, interesować nas będzie jego kwadrat.

Wariancja wyjścia modelu dla danego wzorca wejściowego, wokół rzeczywistej wartości zmiennej objaśnianej, czyli wartość oczekiwana kwadratu odchylenia modelu dla danego wejścia $E((y - f(x, \mathbf{w}))^2 / \mathbf{x})$ z definicji jest równa:

$$E((y - f(\mathbf{x}, \mathbf{w}))^2 / \mathbf{x}) = \int_{-\infty}^{\infty} (y - f(\mathbf{x}, \mathbf{w}))^2 \, p(y / \mathbf{x}) dy \qquad (3.1.14)$$

Zauważmy, że wyrażenie (3.1.14) jest bardzo zbliżone do wyrażenia (3.1.5). W (3.1.5) występuje jeszcze całka ze względu na $\mathbf{x}$, ponieważ interesowała nas tam wartość oczekiwana błędu dla całego zbioru treningowego, a nie dla pojedynczego wzorca wejściowego. Postępując więc niemal dokładnie w taki sam sposób jak przy przekształcaniu (3.1.5) do postaci (3.1.12), otrzymujemy:

$$E((y - f(\mathbf{x}, \mathbf{w}))^2 / \mathbf{x}) = (E(y / \mathbf{x}) - f(\mathbf{x}, \mathbf{w}))^2 + E((y - E(y / \mathbf{x}))^2 / \mathbf{x}) \qquad (3.1.15)$$

W praktyce należy jednak pamiętać o tym, że (3.1.15) stanowi zależność graniczną dla nieskończonego zbioru uczącego. Natomiast zbiór treningowy D jest jedynie pewną próbą, wylosowaną z populacji generalnej definiującej zależność stochastyczną między zmienną objaśnianą y a zmiennymi objaśniającymi $\mathbf{x}$. Jeżeli z tej samej populacji, definiującej tę samą zależność, wylosujemy inny N-elementowy zbiór treningowy, to otrzymamy nieco inny zestaw wag (parametrów) sieci neuronowej (neuronowo-rozmytej). Analizując niepewność modelu, musimy uwzględnić ten efekt.

Powinniśmy więc zmodyfikować nasze pytanie i zastanowić się, jaka będzie wartość oczekiwana błędu dla dowolnego zbioru treningowego, który może zostać wybrany z interesującej nas zależności – oznaczmy ją jako $E_D(E((y - f(x, \mathbf{w}))^2 / \mathbf{x}))$.

$$E_D(E((y-f(\mathbf{x},\mathbf{w}))^2/\mathbf{x})) = \\ = E_D((E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))^2)+E_D(E((y-E(y/\mathbf{x}))^2/\mathbf{x})) \tag{3.1.16}$$

Zauważmy, że drugi człon (3.1.16), jako niezależny od wyjścia modelu, jest również niezależny (stały) względem wybranego zbioru treningowego D. Wynika on, jak to już wcześniej wspominaliśmy, z nieredukowalnego szumu losowego związanego z samym charakterem zależności między zmiennymi y i $\mathbf{x}$. Zależność ta ma charakter stochastyczny, a więc nie potrafimy wyjaśnić, dlaczego dla tej samej wartości $\mathbf{x}$ faktyczne wartości y mogą się między sobą nieco różnić. Człon ten dalej określać będziemy jako wariancję czynnika losowego.

Wartość oczekiwana stałej równa jest tej stałej, a więc mamy:

$$E_D(E((y-f(\mathbf{x},\mathbf{w}))^2/\mathbf{x})) = \\ = E_D((E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))^2)+E((y-E(y/\mathbf{x}))^2/\mathbf{x}) \tag{3.1.17}$$

Pierwszemu członowi (3.1.17) musimy przyjrzeć się dokładniej. Określa on wartość oczekiwaną błędu między faktycznym wyjściem modelu a tym, co powinniśmy otrzymać, czyli funkcją regresji zmiennej y dla danego $\mathbf{x}$. Gdyby funkcje realizowane przez sieci neuronowe (neuronowo-rozmyte) otrzymywane dla różnych zbiorów treningowych były doskonałymi predyktorami wartości oczekiwanej, ten błąd byłby równy 0.

Pierwszy człon (3.1.17) zmodyfikujemy, korzystając z podobnego chwytu jak w przypadku (3.1.5). W pierwszym kroku dodajmy i odejmijmy wartość oczekiwaną wyjścia modelu dla wszystkich możliwych realizacji zbioru treningowego:

$$E_D\left((E(y/\mathbf{x})-f(\mathbf{x},\mathbf{w}))^2\right) = \\ = E_D\left(E(y/\mathbf{x})-E_D(f(\mathbf{x},\mathbf{w}))+E_D(f(\mathbf{x},\mathbf{w}))-f(\mathbf{x},\mathbf{w}))^2\right) \tag{3.1.18}$$

Skorzystajmy ze wzoru na kwadrat sumy oraz addytywności wartości oczekiwanej zmiennej losowej:

$$E_D\left((f(\mathbf{x},\mathbf{w})-E(y/\mathbf{x}))^2\right) = \\ = E_D\left((f(\mathbf{x},\mathbf{w})-E_D(f(\mathbf{x},\mathbf{w})))^2\right)+E_D\left((E_D(f(\mathbf{x},\mathbf{w}))-E(y/\mathbf{x}))^2\right)+ \\ +E_D\left(2(f(\mathbf{x},\mathbf{w})-E_D(f(\mathbf{x},\mathbf{w})))(E_D(f(\mathbf{x},\mathbf{w}))-E(y/\mathbf{x}))\right) \tag{3.1.19}$$

Ponownie łatwo zauważyć, że ponieważ wyrażenie $E_D(f(x, \mathbf{w})) - E(y / \mathbf{x})$ jest stałe względem możliwych zbiorów treningowych D, to – korzystamy

z właściwości wartości oczekiwanej – człon w ostatnim wierszu (3.1.19) jest równy 0:

$$\begin{aligned} & E_D\left(2\left(f(\mathbf{x},\mathbf{w})-E_D(f(\mathbf{x},\mathbf{w}))\right)\left(E_D(f(\mathbf{x},\mathbf{w}))-E(y/\mathbf{x})\right)\right)= \\ & \quad = 2\left(E_D(f(\mathbf{x},\mathbf{w}))-E(y/\mathbf{x})\right)E_D\left(\left(f(\mathbf{x},\mathbf{w})-E_D(f(\mathbf{x},\mathbf{w}))\right)\right)= \\ & \quad = 2\left(E_D(f(\mathbf{x},\mathbf{w}))-E(y/\mathbf{x})\right)\left(E_D(f(\mathbf{x},\mathbf{w}))-E_D(f(\mathbf{x},\mathbf{w}))\right)= \\ & \quad = 2\left(E_D(f(\mathbf{x},\mathbf{w}))-E(y/\mathbf{x})\right)\cdot 0 = 0 \end{aligned} \tag{3.1.20}$$

Ostatecznie więc na podstawie zależności (3.1.19) i (3.1.20) wartość oczekiwaną błędu między faktycznym wyjściem modelu $f(x, \mathbf{w})$ a wartością oczekiwaną zmiennej objaśnianej, dla danego $\mathbf{x}$, $E(y / \mathbf{x})$, możemy zapisać:

$$\begin{aligned} & E_D((f(\mathbf{x},\mathbf{w})-E(y/\mathbf{x}))^2)= \\ & \quad = \underbrace{(E_D(f(\mathbf{x},\mathbf{w}))-E(y/\mathbf{x}))^2}_{\text{obciążenie}} + \underbrace{E_D((f(\mathbf{x},\mathbf{w})-E_D(f(\mathbf{x},\mathbf{w})))^2)}_{\text{wariancja}} \end{aligned} \tag{3.1.21}$$

Analizując problem, dlaczego sieć neuronowa (lub neuronowo-rozmyta) nie stanowi dokładnego predyktora wartości oczekiwanej zmiennej objaśnianej, dla danego wejścia modelu, wykryliśmy więc dwie podstawowe przyczyny tego zjawiska. Pierwsza z nich to fakt, że być może tworzone sieci (systemy rozmyte), średnio rzecz biorąc, nie są równe funkcji regresji. W przypadku wystąpienia tego typu zjawiska w statystyce określa się je jako **obciążenie** modelu. Drugą przyczyną może być fakt, że sieci są jednak czułe na konkretne zbiory danych uczących wykorzystane do ich budowy. Ponieważ uzyskanie perfekcyjnej, idealnej generalizacji modelu neuronowego czy też neuronowo-rozmytego jest w zasadzie niemożliwe, to sieci wytrenowane na różnych zbiorach treningowych będą posiadać nieco odmienne zestawy parametrów wagowych, a co za tym idzie – dawać nieco odmienne wyniki działania. To zjawisko z kolei nazywamy **wariancją** modelu (czasami określa się ją jako wariancję modelu z parametrów, aby odróżnić pojęcie całościowej wariancji wyjściowej modelu i rozważanej tutaj wariancji częściowej, wynikającej ze skończonego charakteru zbioru treningowego).

Wracając więc do wartości oczekiwanej kwadratu odchylenia modelu dla danego wejścia $\mathbf{x}$ (wariancji wyjściowej modelu), danej przez zależność (3.1.17), musimy jeszcze uwzględnić wariancję czynnika losowego (szumu):

$$E_D(E((y - f(\mathbf{x},\mathbf{w}))^2 / \mathbf{x})) = \tag{3.1.22}$$

$$= \underbrace{(E_D(f(\mathbf{x},\mathbf{w})) - E(y/\mathbf{x}))^2}_{\text{obciążenie}} + \underbrace{E_D((f(\mathbf{x},\mathbf{w}) - E_D(f(\mathbf{x},\mathbf{w})))^2)}_{\text{wariancja}} + \underbrace{E((y - E(y/\mathbf{x}))^2 / \mathbf{x})}_{\text{szum}}$$

Uporządkujmy więc trochę nasze rozważania. Analiza źródeł błędu działania modelu prognostycznego, w przypadku konkretnego wejścia prognozy, doprowadziła nas do trzech głównych źródeł jego niepewności: obciążenia modelu, wariancji modelu i wariancji czynnika losowego (szumu). W praktyce w wielu przypadkach musimy jeszcze rozważyć czwartą grupę czynników – niepewność (błędy) zmiennych wejściowych. Scharakteryzujmy pokrótce każde z tych źródeł w kontekście modeli neuronowych i neuronowo-rozmytych.

1. **Obciążenie modelu**. Występuje ono w tych regionach przestrzeni wejściowej, w których wyjście sieci nie jest równe wartości warunkowej oczekiwanej zmiennej objaśnianej y, dla danego wejścia $\mathbf{x}$ (nawet jeśli będziemy rozbudowywać zbiór danych treningowych w nieskończoność). Jeżeli nowy wzorzec wejściowy, dla którego sporządzamy prognozę, wpada w region charakteryzujący się wysokim obciążeniem, wyjście sieci będzie błędne. Możemy wyróżnić dwie podstawowe przyczyny powstawania obciążenia:

a) wynikające z danych: dotyczy to przypadków, gdy dane treningowe nie reprezentują z dostateczną dokładnością populacji generalnej lub zbiór uczący jest poprawnie skonstruowany, ale sieć neuronowa (neuronowo-rozmyta) jest zbyt prosta dla reprezentacji całego zakresu danych wejściowych, jakie w nim występują; z tych powodów w przestrzeni wejść powstawać mogą pewne regiony, w których uformowane odwzorowanie sieci jest zbyt uproszczone i w związku z tym prognoza nie jest równa funkcji regresji; zwykle udział tego elementu w błędzie predykcji redukujemy, stosując standardowe techniki przygotowania danych oraz oceny modelu, takie jak analiza reszt, weryfikacja na zbiorze testowym itp.,

b) wynikające z uczenia: występuje ono, gdy algorytm treningu sieci neuronowej (neuronowo-rozmytej) nie jest w stanie nauczyć jej warunkowej średniej danych, pomimo że model ma dostatecznie bogatą strukturę, aby osiągnąć rozwiązanie; może to być spowodowane błędnie określonym warunkiem zakończenia uczenia (zakończeniem zbyt wczesnym lub zbyt późnym); tym składnikiem błędu zajmujemy się, stosując różne techniki oceny generalizacji (np. walidacja krzyżowa, techniki oparte na wielokrotnym uczeniu, bootstrap itp.).

2. **Wariancja modelu (z wag – parametrów)**. W trakcie każdego procesu uczenia sieci mamy do czynienia z realizacją wielu zmiennych losowych wynikających z pewnych wyborów, jakie muszą zostać dokonane w trakcie jego prowadzenia. Dotyczy to zarówno zbioru danych, jak i samego procesu uczenia:

a) wynikająca z danych: dane wchodzące w skład zbioru treningowego stanowią skończoną realizację pewnego procesu losowego (w statystyce realizacja ta określana jest jako próba); wagi nauczonej sieci neuronowej są więc również zmienną losową; jeśli użyjemy różnych zbiorów danych próbkowanych z tej samej populacji generalnej, w wyniku uczenia otrzymamy różne zestawy wag (parametrów) modelu i w konsekwencji różne prognozy; zwróćmy przy tym uwagę, że losowość ta nie wynika z losowego charakteru inicjalizacji współczynników wagowych; ten element błędu definiowany jest w stosunku do tych samych ustalonych początkowych wartości wag dla każdej próby treningowej,

b) wynikająca z uczenia: wagi (parametry) nauczonej sieci stanowią zmienną losową również z powodu ich losowej inicjalizacji; niezależnie od wyboru metody uczenia sieci neuronowej czy też neuronowo-rozmytej algorytmy treningowe osiągają minima lokalne; ponieważ zazwyczaj funkcja błędu w przestrzeni parametrów ma wiele minimów lokalnych, proces treningu może doprowadzić do różnych efektów, w zależności od wyboru początkowego zestawu wartości wag.

3. **Wariancja czynnika losowego (szumu losowego)**. Dla danego zestawu zmiennych niezależnych (wejść sieci) istnieje nieredukowalny błąd w prognozie zmiennej zależnej (wyjściowej) spowodowany jej losowym charakterem (warunkowa wariancja w rozkładzie wynikowym danych). W teorii liniowych modeli statystycznych wariancja wynikająca z tego komponentu traktowana jest jako stała i szacowana poprzez wariancję reszt modelu na zbiorze treningowym. Należy jednak zwrócić uwagę, że w przypadku modeli nieliniowych założenie o stałości wariancji błędu losowego jest często niespełnione (właściwość tę nazywamy heteroskedastycznością). Uwzględnienie jej we wnioskowaniu o rozkładzie prognozy wymaga wówczas dostarczenia dodatkowego jej estymatora.

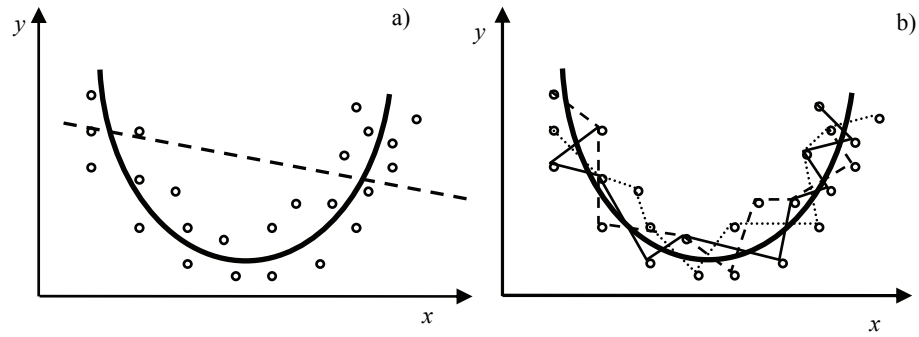
4. **Niepewność (szum) wejściowy**. W przypadku modeli prognostycznych zakładamy zwykle, że zmienne wejściowe (objaśniające) mają charakter deterministyczny i nie są zmiennymi losowymi. Jeśli istnieje niepewność związana z wartościami zmiennych wejściowych (spowodowana np. błędami pomiarowymi, wykorzystaniem jako wejść prognoz i oszacowań statystycznych itp.), to jest ona propagowana poprzez model i traktowana jako część błędu losowego. Istnieje jednak również możliwość wyizolowania tego komponentu i oszacowania go w sposób niezależny.

3.1.3. Wymienność między obciążeniem a wariancją

Analizując czynniki wpływające na fakt, że model prognostyczny nie stanowi dokładnego predyktora wartości oczekiwanej zmiennej objaśnianej w rozkładzie warunkowym, dla danej wartości wzorca wejściowego zmiennych objaśniających, wyodrębniliśmy wśród nich dwa główne komponenty, mianowi-

cie obciążenie modelu oraz jego wariancję (patrz zależność (3.1.21)). Te dwa źródła błędu modeli analizy danych mają przy tym charakter współzależny, a nawet więcej, możemy mówić o pewnej wymienności między nimi.

Zasadniczo, spoglądając na problem dopasowania pewnej funkcji do danych z punktu widzenia obciążenia i wariancji, możemy mówić o dwóch przeciwstawnych skrajnościach zaprezentowanych poglądowo na rysunku 3.1.2. W pierwszym przypadku, w części a), przyjmijmy, że dopasowywana do danych funkcja to ustalona prosta nieposiadająca żadnych parametrów wolnych (szacowanych na podstawie danych). Niezależnie więc od wyboru podzbioru danych, prosta położona jest zawsze tak samo – wariancja modelu wynosi więc zero. Trudno jednak uznać, że dopasowana funkcja dobrze odzwierciedla średnie wartości zmiennej y, dla różnych argumentów x. Mamy więc do czynienia z dużym obciążeniem modelu.



Rysunek 3.1.2. Schematyczna ilustracja wymienności między obciążeniem i wariancją modelu
Źródło: opracowanie własne

Drugi skrajny przypadek zaprezentowany został w części b) rysunku 3.1.2. Dopasowywana funkcja uzyskana została poprzez interpolację dla wybranego podzbioru danych, przy użyciu łamanej. Jak widzimy, obciążenie w tym przypadku jest niewielkie. Jeżeli liczba punktów danych będzie rosła do nieskończoności, to łamana będzie coraz dokładniej odzwierciedlać aproksymowaną funkcję. Ponieważ jednak funkcja interpolacyjna musi przechodzić przez definiujące ją punkty, to w zależności od wyboru wykorzystywanego podzbioru danych, otrzymujemy nieco inną łamaną (linia ciągła, wykreskowana, wykropkowana). W związku z tym mamy do czynienia z dużą wariancją modelu zakłócającą poprawne wyniki modelowania.

Jak więc widzimy, między obciążeniem a wariancją modelu istnieje pewna naturalna wymienność. Funkcja ściśle dopasowana do danych będzie miała tendencję do dużej wariancji, a przynajmniej to wariancja będzie miała główny udział w błędzie opartego na niej modelu. Możemy redukować wariancję

poprzez upraszczanie funkcji, ale jeśli posuniemy ten proces za daleko, to wzrosnąć z kolei może obciążenie i tym razem to ono będzie główną przyczyną znacznego błędu (Bishop 1995).

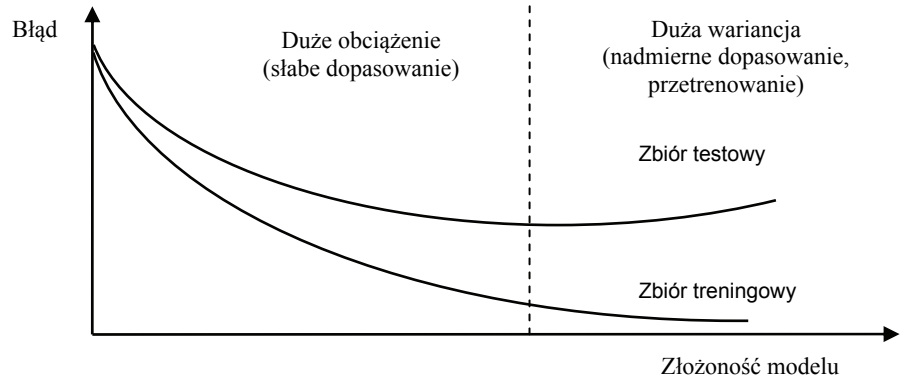
Powstaje w związku z tym pytanie: jak wygląda sytuacja w przypadku modeli neuronowych i systemów rozmytych, których wykorzystanie do prognozy zapotrzebowania na energię elektryczną i moc omawialiśmy w rozdziale 2? W którym punkcie należy je umieścić między skrajnymi przypadkami z rysunku 3.1.2? Otóż zarówno sieci neuronowe, jak i neuronowo-rozmyte należą do kategorii modeli indukcyjnych, potrafiących dobrze dopasowywać się do danych. W przypadku omawianych kategorii modeli istnieje dobrze rozbudowana teoria aproksymacji, przede wszystkim należą one do kategorii tzw. uniwersalnych aproksymatorów (dla sieci MLP patrz Hornik, Stinchcombe, White 1989, dla systemów neuronowo-rozmytych FBF np. Wang, Mendel 1992; Zeng, Singh 1995, dla systemów rozmytych typu Takagi–Sugeno: Ying 1998a, b). Oznacza to, mówiąc ogólnie, że dla dowolnej funkcji (w naszym przypadku możemy ograniczyć się do kategorii funkcji ciągłych) można zbudować model neuronowy, czy też neuronowo-rozmyty, który będzie ją aproksymował z dowolnie dużą dokładnością.

Sieci neuronowe i systemy rozmyte należą więc do kategorii modeli charakteryzujących się niewielkim obciążeniem, natomiast dużą wariancją. Ujmując rzecz dokładniej, na przykładzie procesu uczenia tego typu modeli dostrzec możemy doskonałą ilustrację wspomnianej wcześniej wymienności między obciążeniem a wariancją modelu. Manifestuje się ona w postaci tzw. efektu przetrenowania albo nadmiernego dopasowania modelu do danych. Zjawisko to jest dobrze znane i opisywane w każdym podstawowym podręczniku poświęconym zagadnieniem sieci neuronowych czy też uczenia statystycznego (np. Hertz, Krogh, Palmer 1993; Korbicz, Obuchowicz, Uciński 1994; Masters 1996; Żurada, Barski, Jędruch 1996; Zieliński 2000), a więc w tym miejscu wyjaśnimy tylko przyczyny jego powstawania oraz konsekwencje dla niepewności modelu.

Stosunkowo proste modele, o małej liczbie jednostek w sieci neuronowej czy reguł w systemie rozmytym, mogą okazać się zbyt mało rozbudowane i nie móc osiągnąć właściwej dokładności aproksymacji wartości oczekiwanej zmiennej wyjściowej. Aby rozwiązać ten problem, możemy zwiększać złożoność modelu, dodając nowe neurony (lub ich warstwy) albo reguły rozmyte. Rozbudowywanie struktury modelu przy określonej liczbie wzorców treningowych wymaga jednak pewnej uwagi. Jeżeli model staje się za bardzo złożony w stosunku do rozmiaru zbioru treningowego, to zbyt długa kontynuacja procesu uczenia doprowadza do nadmiernego jego dopasowania do konkretnych danych, co z kolei skutkuje wzrostem błędu z powodu wzrostu wariancji modelu.

Schematyczną ilustrację efektu przetrenowania widzimy na rysunku 3.1.3. Konkretna sieć neuronowa lub system rozmyty z powodu dopasowania do zbioru treningowego nie przybliża właściwej funkcji regresji, tylko dane z tego

konkretnego zestawu wykorzystanego do uczenia. Obciążenie jest małe, tzn. dla rosnącej liczby wzorców danych uczących aproksymacja byłaby coraz dokładniejsza. Natomiast wzrasta błąd wynikający z wariancji modelu. Jeśli wykonujemy prognozę dla wzorca danych wejściowych, który nie występuje w zbiorze treningowym, to błąd sieci neuronowej (neuronowo-rozmytej) zaczyna rosnąć. Na rysunku 3.1.3 obrazuje to wzrost krzywej błędu na zbiorze testowym.



Rysunek 3.1.3. Schematyczna ilustracja wymienności między obciążeniem i wariancją w procesie dopasowywania modelu do danych
Źródło: opracowanie własne

Czy jednoczesna minimalizacja obciążenia i wariancji modelu jest możliwa? Najlepszym oczywiście sposobem okaże się tutaj wykorzystanie jakiejś dodatkowej wiedzy. Na przykład jeśli wiemy, że zależność między zmienną wejściową $\mathbf{x}$ a wyjściową y ma charakter liniowy, to zastosowanie w systemie prognostycznym modelu liniowego, zamiast, powiedzmy, sieci neuronowej, powinno dać dobre efekty – mniejszą wariancję, ponieważ model jest prostszy oraz ma mniej parametrów, natomiast obciążenie nie powinno wzrosnąć. Stanie się tak oczywiście pod warunkiem, że nasza wiedza na temat liniowego charakteru zależności między zmiennymi jest poprawna.

W procesie tworzenia systemu prognostycznego opartego na nieliniowych złożonych modelach aproksymacyjnych, takich jak sieci neuronowe czy systemy rozmyte, znalezienie jakiegoś uporządkowanego podejścia do jednoczesnej minimalizacji obciążenia i wariancji jest, niestety, dosyć trudne. Istnieją pewne metody konstruowania optymalnej struktury modelu, ale przy zazwyczaj ograniczonych ilościach danych historycznych i dużych nakładach obliczeniowych potrzebnych do ich realizacji (zwłaszcza w przypadku sieci neuronowych), rozwiązania te mają znaczenie raczej teoretyczne.

W odniesieniu do omawianych modeli niezbędne jest więc znalezienie pewnej równowagi pomiędzy obciążeniem a wariancją. Polega to zazwyczaj na

tym, że podczas uczenia modelu staramy się wychwycić punkt, w którym błąd wynikający z obciążenia jest już mały, a błąd wynikający z wariancji jeszcze nie zaczął rosnąć, oceniając model nie tylko w kategoriach dopasowania do danych treningowych, ale również w kategoriach generalizacji na dodatkowym zbiorze testowym, tzw. zbiorze walidacyjnym. Tę dobrze znaną i najczęściej chyba obecnie wykorzystywaną metodę określa się mianem walidacji krzyżowej (*cross-validation*).

Jakie są konsekwencje takiego sposobu postępowania dla niepewności modelu i sposobu jej modelowania? Aby zminimalizować ryzyko wystąpienia obciążenia z powodu zbyt prostej struktury modelu, jego ocenę wykonuje się zazwyczaj dla kilku wariantów sieci neuronowej (systemu neuronowo-rozmytego). Jeżeli ponadto dane do budowy systemu zostały przygotowane poprawnie, to komponent obciążenia wynikający z doboru danych również powinien być minimalny. Jeżeli zarówno zbiór treningowy, jak i testowy są reprezentatywne w całej przestrzeni wejść systemu, to ponieważ wybierane były z tej samej populacji ogólnej, różnica w błędzie modelu dla obu zbiorów powinna wynikać przede wszystkim z wariancji. Obciążenie powinno być niewielkie.

Podsumowując więc, obciążenie poprawnie, zgodnie ze wszystkimi kanonami sztuki, przygotowanego modelu, ocenianego na podstawie jakości generalizacji metodą walidacji krzyżowej, powinno być nieznaczne, przynajmniej w porównaniu z jego wariancją. Modele nieliniowe o bogatych możliwościach aproksymacyjnych, takie jak sieci neuronowe lub systemy rozmyte, należą do kategorii systemów o niskim obciążeniu i wysokiej wariancji. W związku z tym w dalszej części naszej pracy zakładać będziemy, że wykorzystywane modele prognostyczne są nieobciążone, a źródłami ich błędu są przede wszystkim wariancja wynikająca z parametrów, wariancja czynnika losowego i (jeżeli występuje) niepewność (wariancja) zmiennych wejściowych.

3.2. Charakterystyka rozkładu prognozy

3.2.1. Warunkowy rozkład prawdopodobieństwa prognozowanego zjawiska

Prognozowanie stanowi pomocniczą funkcję zarządzania. Nie jest ono celem samym w sobie, lecz ważnym narzędziem pozwalającym zredukować naszą niepewność co do istotnych zjawisk wpływających na podejmowane decyzje. Jak wspomnieliśmy na wstępie bieżącego rozdziału, możemy mieć czasami do czynienia z sytuacją, kiedy właściwe podejście do pewnych decyzji, oszacowanie elementów ryzyka poszczególnych skutków z nich wynikających, będzie

wymagać informacji otrzymywanych z całego wynikowego warunkowego rozkładu prawdopodobieństwa prognozowanego zjawiska.

Istnieją oczywiście metody analizy danych, które pozwalają na bezpośrednią estymację funkcji gęstości rozkładu prognozowanej zmiennej $p(y / \mathbf{x})$. Są one jednak stosunkowo rzadko wykorzystywane w praktyce. Zazwyczaj wymagają dużych ilości danych oraz dużych nakładów obliczeniowych do budowy modelu. W większości są to ciągle metody laboratoryjne, o słabo jeszcze poznanych właściwościach. W obecnej pracy idziemy niejako w drugą stronę. Mając stosunkowo dobrze już ugruntowane takie technologie jak sieci neuronowe czy systemy rozmyte, wchodzące w mniejszym lub większym stopniu do katalogu praktycznych narzędzi prognostycznych, określimy sposoby otrzymania na ich podstawie gęstości prawdopodobieństwa prognozowanej zmiennej $p(y / \mathbf{x})$.

Spróbujmy więc podsumować, co wiemy na temat rozkładu prognozowanej zmiennej dla danego wejścia modelu. Niepewność opisywana przez ten rozkład prawdopodobieństwa jest efektem niepewności wyjścia modelu, a więc zasadniczo rozkład ten będzie definiowany przez rozkład wyjścia modelu. Przede wszystkim, jak to przedstawialiśmy w punkcie 3.1.l, wyjście nauczonego modelu $f(\mathbf{x}, \mathbf{w})$, czyli uzyskana za jego pomocą prognoza, stanowi oszacowanie wartości oczekiwanej $E(y / \mathbf{x})$ (patrz zależność (3.1.13)). Innymi słowy, wyjście modelu prognozy zapotrzebowania na energię elektryczną (lub moc) daje nam najbardziej prawdopodobną, średnią wartość tego zapotrzebowania y, jakiej spodziewamy się dla danego zestawu danych wejściowych $\mathbf{x}$.

Potrafimy również nieco powiedzieć o rozproszeniu prawdopodobieństwa w rozkładzie $p(y / \mathbf{x})$, czyli o wariancji tego rozkładu. Wariancja prognozowanej zmiennej równa będzie wariancji wyjścia modelu dla danego wzorca wejściowego $\mathbf{x}$, którą oznaczymy $\sigma^2_y(\mathbf{x})$. W ogólnym przypadku na podstawie (3.1.22) wiemy, że niepewność wyjścia określana jest przez komponenty obciążenia, wariancji modelu wynikającej z niepewności jego parametrów (wag) i wariancji czynnika losowego. Dyskutując w punkcie 3.1.3 problem wymienności między obciążeniem a wariancją, wskazaliśmy, że w przypadku poprawnie zbudowanego modelu neuronowego lub neuronowo-rozmytego obciążenie modelu w porównaniu z jego wariancją z parametrów jest niewielkie i można je zaniedbać. Zakładać zatem będziemy, że **model jest nieobciążony**.

Przyjmując powyższe założenie, na podstawie (3.1.22) możemy powiedzieć, że wariancja rozkładu prawdopodobieństwa prognozowanej zmiennej $\sigma^2_y(\mathbf{x})$, dla danego wzorca danych wejściowych $\mathbf{x}$, może zostać oszacowana przez:

$$\begin{aligned}\sigma_y^2(\mathbf{x}) &= E_D\left(E\left((y - f(\mathbf{x},\mathbf{w}))^2 / \mathbf{x}\right)\right) = \\ &= E_D\left(\left(f(\mathbf{x},\mathbf{w}) - E_D(f(\mathbf{x},\mathbf{w}))\right)^2\right) + E\left((y - E(y/\mathbf{x}))^2 / \mathbf{x}\right) = \\ &= \sigma_{\mathbf{w}}^2(\mathbf{x}) + \sigma_{\varepsilon}^2(\mathbf{x})\end{aligned} \tag{3.2.1}$$

gdzie $\sigma^2_{\mathbf{w}}(\mathbf{x})$ oznacza wariancję wyjścia modelu prognostycznego wynikającą z niepewności parametrów (wag), zaś $\sigma^2_{\varepsilon}(\mathbf{x})$ wariancję czynnika losowego (szum losowy). Metody wyznaczania tych dwóch komponentów omawiać będziemy, odpowiednio, w podrozdziałach 3.3 oraz 3.4.

Pamiętać również należy, że jeżeli z jakiegoś powodu podczas eksploatacji modelu prognostycznego występuje niepewność podawanych wejść (np. jako wartości wejściowe wykorzystywane są także prognozy albo wartości obciążone błędami pomiarowymi), to szacując wariancję rozkładu prawdopodobieństwa prognozowanej zmiennej, należy uwzględnić komponent wynikający z tego faktu:

$$\sigma_y^2(\mathbf{x}) = \sigma_{\mathbf{w}}^2(\mathbf{x}) + \sigma_{\varepsilon}^2(\mathbf{x}) + \sigma_{\mathbf{x}}^2(\mathbf{x}) \tag{3.2.2}$$

gdzie $\sigma^2_{\mathbf{x}}(\mathbf{x})$ oznacza wariancję wyjścia modelu spowodowaną niepewnością wejść. Problem ten będziemy prezentować w podrozdziale 3.5.

3.2.2. Przedziały prognozy

Problematyką wykorzystania oszacowań rozkładu prognozowanego krótkoterminowego zapotrzebowania na energię elektryczną w konkretnych zagadnieniach decyzyjnych zajmiemy się w rozdziale 4. W rozdziale bieżącym uzyskane wyniki oceniać będziemy głównie pod kątem probabilistycznym. Dokładniej rzecz biorąc, wątkiem przewodnim pozwalającym na analizę i porównanie metod oszacowania warunkowego rozkładu prawdopodobieństwa wyjścia modelu dla danego wejścia będą teraz tzw. przedziały prognozy.

Przedziałem prognozy z prawdopodobieństwem α (α-przedziałem prognozy) nazywamy taki przedział wartości prognozowanej zmiennej $[d_{y/\mathbf{x}}(\alpha), g_{y/\mathbf{x}}(\alpha)]$, w którym powinna się ona znaleźć z prawdopodobieństwem α:

$$\Pr_{y/\mathbf{x}}(\, y \in [d_{y/\mathbf{x}}(\alpha), g_{y/\mathbf{x}}(\alpha)]\,) = \alpha \tag{3.2.3}$$

Dla danej prognozy istnieje oczywiście nieskończenie wiele przedziałów spełniających warunek (3.2.3). Najczęściej więc przyjmuje się symetryczny charakter przedziału prognozy, tzn. taki, dla którego prawdopodobieństwo, że wartość prognozowanej zmiennej znajdzie się jednak poniżej $d_{y/\mathbf{x}}(\alpha)$ jest takie samo jak prawdopodobieństwo tego, że przekroczy ona $g_{y/\mathbf{x}}(\alpha)$, i wynosi ono

$(1 - \alpha)/2$. W dalszej części książki, mówiąc o przedziałach prognozy, zakładać będziemy symetryczny ich charakter.

Przedziały te mają więc istotne znaczenie dla decydenta, który wykorzystuje prognozę do oszacowania wartości rozważanych alternatyw decyzyjnych. Pozwalają one stworzyć pewne ramy, w których powinna mieścić się prognoza z przyjętym (zazwyczaj rozsądnie dużym) prawdopodobieństwem, oraz analizować poziom bezpieczeństwa i stabilności wybieranych opcji rozwiązania problemu. Tym niemniej samo ich określanie odbywa się na zasadach czysto probabilistycznych, niepowiązanych z żadnym konkretnym problemem decyzyjnym.

Do wyznaczenia przedziału dla konkretnej prognozy musimy określić warunkowy rozkład prawdopodobieństwa prognozowanej zmiennej y dla danego wzorca wejściowego modelu $\mathbf{x}$, np. jego funkcję gęstości $p(y / \mathbf{x})$ albo dystrybuantę $P(y / \mathbf{x})$. Przyjmując symetryczny charakter przedziału prognozy, jego krańce możemy wyznaczyć na podstawie zależności:

$$\Pr_{y/\mathbf{x}}(y < d_{y/\mathbf{x}}(\alpha)) = (1 - \alpha)/2 \tag{3.2.4a}$$

$$\Pr_{y/\mathbf{x}}(y > g_{y/\mathbf{x}}(\alpha)) = (1 - \alpha)/2 \tag{3.2.4b}$$

Zauważmy, że w przypadku pierwszej równości (3.2.4a) dla dolnego krańca przedziału prognozy możemy ją zapisać za pomocą dystrybuanty rozkładu warunkowego zmiennej y: $P(d_{y/\mathbf{x}}(\alpha) / \mathbf{x}) = (1 - \alpha)/2$. W przypadku warunku (3.2.4b) zauważmy, że $\Pr_{y/\mathbf{x}}(y > g_{y/\mathbf{x}}(\alpha)) = 1 - \Pr_{y/\mathbf{x}}(y \le g_{y/\mathbf{x}}(\alpha))$. Warunek ten możemy więc również zapisać przy użyciu dystrybuanty: $P(g_{y/\mathbf{x}}(\alpha) / \mathbf{x}) = 1 - (1 - \alpha)/2 = (1+\alpha)/2$. Ostatecznie więc dolny i górny kraniec przedziału prognozy wyznaczymy za pomocą zależności:

$$d_{y/\mathbf{x}}(\alpha) = Q_{y/\mathbf{x}}((1 - \alpha)/2) \tag{3.2.5a}$$

$$g_{y/\mathbf{x}}(\alpha) = Q_{y/\mathbf{x}}((1 + \alpha)/2) \tag{3.2.5b}$$

gdzie $Q_{y/\mathbf{x}}(\alpha)$ jest kwantylem dla prawdopodobieństwa α, czyli α-kwantylem w rozkładzie warunkowym prawdopodobieństwa zmiennej prognozowanej y dla danego wzorca wejściowego $\mathbf{x}$.

Zwróćmy jeszcze uwagę na fakt, że wyznaczone w praktyce przedziały prognozy stanowić będą pewne oszacowania faktycznych wartości. Niezbędna więc może być jakaś forma ich weryfikacji empirycznej poprzez przetestowanie i analizę uzyskanych wyników dla pewnego zbioru danych. Zastanówmy się teraz nad interpretacją wyników uzyskiwanych w procesie takiego testowania (Masters 1995).

Przyjmijmy, że przedział prognozy przetestowany został na zbiorze m obserwacji, przy czym okazało się, że w k przypadkach rzeczywista wartość

zmiennej y znalazła się poza nim. Zakładać przy tym będziemy, że każda prognoza testowa była wykonywana niezależnie od innych. Oznaczmy przez $p = (1 - \alpha)$ prawdopodobieństwo wyjścia wartości rzeczywistej prognozowanej zmiennej poza przedział prognozy. Zauważmy, że prawdopodobieństwo p dotyczy wyniku pojedynczego testowania modelu. Przy m-elementowym zbiorze testowym mamy do czynienia z m-krotnym powtórzeniem niezależnych prób, w którym odniesiono k „sukcesów", czy też właściwie w tym przypadku słuszniej byłoby powiedzieć „niepowodzeń". Mamy więc oczywiście do czynienia ze znanym z podstaw rachunku prawdopodobieństwa schematem Bernoulliego, w którym prawdopodobieństwo liczby sukcesów podlega rozkładowi dwumianowemu, którego funkcja prawdopodobieństwa dana jest wzorem:

$$B(k,m,p) = \binom{m}{k} p^k (1-p)^{m-k} = \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k} \tag{3.2.6}$$

Zauważmy przy tym, że liczba przypadków przekroczenia granic przedziału prognozy podlega rozkładowi dwumianowemu niezależnie od kształtu prognozowanego rozkładu warunkowego zmiennej y, dla którego przedział ten został wyznaczony.

Do różnego typu procesów wnioskowania i oceny wybranych wartości k lub p raczej korzystać będziemy, rzecz jasna, z dystrybuanty rozkładu dwumianowego. Dla przykładu, jeśli podejrzewamy, że wynik testowania k sugeruje błędne określenie krańców przedziału prognozy i konieczność korekty faktycznego prawdopodobieństwa przedziału na wartość p_1, to aby oszacować prawdopodobieństwo takiego faktu musimy przeanalizować wartość dystrybuanty:

$$\sum_{i=0}^{k} \frac{m!}{i!(m-i)!} p_1^i (1-p_1)^{m-i} \tag{3.2.7}$$

3.2.3. Nieparametryczne i parametryczne podejście do oszacowania rozkładu prognozy

W punkcie 3.2.1 zastanawialiśmy się nad możliwymi źródłami informacji, które mogą posłużyć do określenia wartości oczekiwanej i wariancji, czyli pewnych parametrów rozkładu wyjścia modelu prognostycznego dla danego wzorca wejściowego. Powstaje jednak pytanie, w jaki sposób uzyskać informacje o całości warunkowego rozkładu prawdopodobieństwa prognozowanej wielkości. Ogólnie rzecz biorąc, możliwe są tutaj w zasadzie dwa podstawowe podejścia: parametryczne i nieparametryczne.

Często potrafimy dobrać pewien określony model błędu prognozy i, w konsekwencji, określić kształt rozkładu prawdopodobieństwa prognozowanej wielkości. Wówczas w przypadku konkretnej prognozy wystarczy oszacować parametry tego rozkładu dla danego wzorca wejściowego modelu. Mówimy wtedy o parametrycznym oszacowaniu (estymacji) rozkładu prognozy. Jeżeli więc wiemy, w jaki sposób wykorzystać źródła informacji omawiane w punkcie 3.2.1, czyli dla danego typu modelu potrafimy wyznaczyć wartość oczekiwaną i wariancję jego wyjścia, a ponadto dla wykorzystywanego modelu spełnione są pewne założenia związane z tymi obliczeniami, to w praktyce możemy określić cały rozkład prognozowanej wielkości.

Możliwa jest jednak sytuacja, w której jakiś z wymienionych tu warunków nie jest spełniony. Nie potrafimy określić rozkładu błędu lub dla danego typu modelu nie znamy metody oszacowania, zwłaszcza wariancji wyjściowej. Może się zdarzyć, że decydent (użytkownik) nie zna dokładnie struktury gotowego systemu prognostycznego, z którego korzysta, a projektant nie przewidział konieczności określania przez system wariancji wyjściowej modelu dla danej prognozy. W takim przypadku trudno przecież mówić o analitycznym oszacowaniu działania modelu o nieznanej strukturze. W końcu może się również zdarzyć, że dla konkretnej sytuacji nie są spełnione założenia związane z wnioskowaniem o parametrach rozkładu wyjściowego dla modeli tego typu. We wszystkich tego rodzaju przypadkach właściwie nie mamy innego wyjścia, jak posłużyć się podejściem nieparametrycznym i próbować oszacować rozkład prognozowanej wielkości na podstawie testowania modelu.

Podejście oparte na rozkładzie empirycznym jest w zasadzie dosyć proste i intuicyjne. Polega ono na wykorzystaniu do analizy i oszacowania niepewności otrzymanej prognozy po prostu bezpośrednio zbioru pomiarów jej błędów, co wydaje się rozsądnym rozwiązaniem. Procedurę ich porządkowania i wykorzystania możemy przedstawić następująco (Masters 1995).

Przyjmijmy, że dany jest pewien zbiór danych testowych, zawierający obserwacje złożone z wzorców wejściowych prognozy oraz odpowiadających im rzeczywistych wartości zmiennej wyjściowej:

$$T = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}, k = 1, \ldots, M \tag{3.2.8}$$

Istotne jest przy tym, że w zbiorze testowym T muszą znajdować się inne obserwacje niż w zbiorze treningowym D, ponieważ interesuje nas analiza błędu generalizacji modelu (jego działanie w nowych sytuacjach), a nie jego dopasowania do danych uczących. Model testujemy na zbiorze T, wyznaczając M odchyleń (błędów) prognozy dla poszczególnych wzorców testowych:

$$e_k = f(\mathbf{x}_k, \mathbf{w}) - y_k, k = 1, \ldots, M \tag{3.2.9}$$

W kolejnym kroku odchylenia e_k porządkowane są w sposób niemalejący, dając zbiór statystyk porządkowych rozkładu błędu $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(M)}$. Pamiętać należy, że w przeciwieństwie do oceny dokładności działania modelu dla całego zbioru danych (błędów średnich), w tym przypadku statystyki $e_{(k)}$ muszą zachować znak wykorzystywanych odchyleń, a nie tylko ich wielkość.

Dystrybuantę empiryczną rozkładu błędu dla naszej M-elementowej próby T możemy zapisać następująco:

$$P_M(e) = \begin{cases} 0 & , e < e_{(1)} \\ \dfrac{k}{M} & , e_{(k)} \leq e < e_{(k+1)} \\ 1 & , e_{(M)} \leq e \end{cases} \tag{3.2.10}$$

Innymi słowy, $P_M(e)$ jest równe części zbioru błędów, które są mniejsze bądź równe e. Dystrybuantę empiryczną warunkowego rozkładu prawdopodobieństwa prognozowanej wielkości możemy zapisać wówczas za pomocą wzoru:

$$P(y / \mathbf{x}) = P_M(f(\mathbf{x}, \mathbf{w}) - y) \tag{3.2.11}$$

Dystrybuantę empiryczną błędu $P_M(e)$ i zbiór statystyk porządkowych rozkładu błędu $e_{(k)}$, $k = 1, \ldots,$ M możemy wykorzystać, dla przykładu, do wyznaczenia przedziału prognozy błędu. Jeżeli jego prawdopodobieństwo ma wynosić α, to prawdopodobieństwo wyjścia poza każdy z jego krańców wynosi $p = (1 - \alpha)/2$. Przedział prognozy błędu $[d_e(\alpha), g_e(\alpha)]$ wyznaczony będzie z obu stron przez dwie skrajne statystyki porządkowe błędu $e_{(d)}$, $e_{(g)}$ spełniające warunki:

$$P_M(e_{(d)}) = p \text{ oraz } P_M(e_{(g)}) = 1 - p. \tag{3.2.12}$$

Wyznaczenie $e_{(d)}$, $e_{(g)}$ stanowi prostą operację. Wystarczy ze zbioru statystyk porządkowych rozkładu błędu $e_{(k)}$, $k = 1, \ldots, M$ odrzucić z obu krańców po $m = M \cdot p$ elementów skrajnych. Pozostałe na obu krańcach statystyki porządkowe wyznaczać będą wartości $e_{(d)}$, $e_{(g)}$. Jeżeli interesuje nas, przykładowo, przedział błędu dla $\alpha = 90\%$, to musimy z obu stron zbioru odrzucić po 5% błędów.

Po wyznaczeniu $e_{(d)}$, $e_{(g)}$ możemy oczywiście łatwo znaleźć przedział prognozy dla warunkowego rozkładu wyjścia modelu przy danym wzorcu wejściowym (pamiętając, że $e_{(d)}$ powinno być wielkością ujemną):

$$[d_{y/\mathbf{x}}(\alpha), g_{y/\mathbf{x}}(\alpha)] = [f(\mathbf{x}, \mathbf{w}) + e_{(d)}, f(\mathbf{x}, \mathbf{w}) + e_{(g)}] \tag{3.2.13}$$

Zwróćmy uwagę na pewien szczegół. Otóż przybliżamy tutaj ciągły rozkład prognozy za pomocą dyskretnego rozkładu empirycznego błędu. W związku z tym często możemy mieć do czynienia z sytuacją, że $m = M \cdot p$ nie będzie wartością całkowitą. W takim przypadku decyzja o wyborze wartości całkowitej m obarczona jest pewną dowolnością. Przyjmuje się raczej, że powinniśmy przyjąć rozwiązanie bezpieczniejsze, zaokrąglając m w dół, kosztem nawet pewnego przeszacowania wielkości przedziału prognozy. W praktyce jednak w zagadnieniach krótkoterminowego prognozowania zapotrzebowania na energię elektryczną, przy pozostających zazwyczaj do dyspozycji dosyć dużych rozmiarach zbiorów danych, poszczególne statystyki porządkowe rozkładu błędu rozmieszczone są dosyć „gęsto", tak więc problem ten ma na ogół niewielkie znaczenie.

Dalsze bardziej szczegółowe informacje na temat wnioskowania przy wykorzystaniu statystyk porządkowych rozkładu prawdopodobieństwa znaleźć można w specjalistycznej literaturze przedmiotu (np. David, Nagaraja 2003).

Analizując podejście empiryczne do oszacowania rozkładu prawdopodobieństwa prognozowanej wielkości, musimy jednak pamiętać, że opiera się ono na kilku fundamentalnych założeniach (Masters 1995).

1. Staramy się określić rozkład prognozy, próbkując błąd generalizacji modelu. Oszacowanie dystrybuanty rozkładu błędu zależne jest od konkretnego zbioru testowego T, który wykorzystujemy do wyznaczenia próbek błędów, i dla różnych zbiorów może się ono zmieniać. Liczba wykorzystywanych próbek nie może więc być zbyt mała. Zbiór testowy musi być również dokładną i właściwą reprezentacją populacji ogólnej, z którą będziemy mieli do czynienia w praktyce. Wyznaczony rozkład empiryczny odnosił się będzie tylko do populacji ogólnej reprezentowanej przez zbiór testowy.

2. Aby otrzymać próbki błędów, model powinien być przetestowany dla obserwacji, które w żaden sposób nie zostały wykorzystane w procesie jego tworzenia. W punkcie 3.1.3 prezentowaliśmy problem wymienności między obciążeniem modelu a jego wariancją. Złożone nieliniowe modele o bogatych możliwościach aproksymacyjnych mają w procesie uczenia tendencję do nadmiernego dopasowania się do danych. Wykorzystanie w procesie testowania wzorców treningowych spowoduje, że w oszacowaniu rozkładu prognozy nie doszacujemy odpowiednio elementu wariancji modelu spowodowanej niepewnością jego parametrów. W efekcie otrzymany rozkład prognozy może być zbyt optymistyczny co do jego rozproszenia (zbyt „wąski"). Pamiętać należy, że w praktycznym zastosowaniu mała jest możliwość, że model będzie sporządzał prognozy w dokładnie takich samych warunkach, jakie zostały zawarte w danych treningowych.

3. Zakłada się, że poszczególne próbki błędów są niezależne i mają jednakowy rozkład. Jeżeli np. błąd prognozy zależy od jej wartości, może to doprowadzić do powstania pewnych asymetrii rozkładu.

Analizując trzy zaprezentowane założenia z punktu widzenia krótkoterminowego prognozowania zapotrzebowania na energię (moc), należy stwierdzić, że pierwsze dwa z nich rzadko stanowią istotny problem. Nasze wieloletnie doświadczenia wskazują, że zależności między kluczowymi zmiennymi w tej dziedzinie ewoluują raczej w powolnym tempie. Oczywiście nie mówimy tutaj o kategoriach bezwzględnych, gdzie obserwujemy wyraźne zmiany zarówno w wielkości samego zapotrzebowania, jak i przesunięcia w pewnych zjawiskach okresowych: wyrównywanie szczytów i dolin dobowych, przesunięcia obciążenia między sezonem zimowym a letnim itp. Zauważmy jednak, że dynamika zmian tych zjawisk również ma charakter długookresowy, wieloletni. Zależności w kategoriach względnych, np. między zapotrzebowaniem na energię dla danej godziny w dniu dzisiejszym i jutrzejszym wykazują nawet większą stałość.

Fakty te powodują, że na potrzeby krótkoterminowego prognozowania zapotrzebowania na energię możemy swobodnie dysponować dosyć długimi dwu-, trzyletnimi, a nawet nieco dłuższymi zbiorami niemal codziennych obserwacji (pozostawiamy na boku kwestię dni nietypowych, które jak wskazywaliśmy w punkcie 2.2.5 wymagają nieco odmiennego potraktowania). Pozwala to na wybranie dużych, reprezentatywnych zbiorów danych, zarówno do tworzenia, jak i oceny działania modelu, spełniających założenia poczynione w punktach pierwszym i drugim. Nie znaczy to naturalnie, że co pewien czas model nie powinien zostać zaktualizowany i przeuczany dla nowo pojawiających się danych. Tym niemniej w przedziale tego kilkuletniego okresu możemy poruszać się dosyć bezpiecznie.

Niestety w przypadku założenia poczynionego w punkcie trzecim sytuacja ma się nieco odmiennie. W zasadzie mamy tam do czynienia z dwoma założeniami, pierwsze z nich dotyczy niezależności, a drugie jednakowego rozkładu wszystkich próbek błędów wykorzystywanych do oszacowania rozkładu wyjścia modelu. Pierwsze z tych założeń sprawia mniejsze kłopoty. Dla dobrze zbudowanego, nieobciążonego modelu prognostycznego na ogół błędy powinny być niezależne lub niemal niezależne. Ponadto założenie to ma mniejsze znaczenie dla oszacowania rozkładu prognozy jako takiego, a większe na temat pewnych stwierdzeń probabilistycznych, jakie możemy poczynić dla jego parametrów, np. dla przedziałów prognozy. Nie możemy np. użyć do wnioskowania o prawdopodobieństwie granic przedziału prognozy rozkładu dwumianowego (3.2.6), tak jak to sugerowaliśmy w punkcie 3.2.2, który zakłada właśnie niezależność poszczególnych próbek błędów. Istnieją jednak pewne bardziej odporne metody wnioskowania na temat kwantyli rozkładu przy użyciu jego statystyk porządkowych (patrz np. David, Nagaraja 2003), pozostają one jednak poza zakresem tematycznym naszej pracy.

Drugie założenie poczynione w punkcie trzecim dotyczące jednakowego rozkładu wszystkich próbek błędów ma charakter naturalny. Tworzymy przecież jeden empiryczny rozkład błędu, który ma być wykorzystywany dla wszystkich

wykonywanych prognoz. Musimy więc być pewni, że niezależnie od warunków prognozy definiowanych przez jej wzorzec danych wejściowych, rozkład prawdopodobieństwa prognozowanej wielkości wokół jej wartości oczekiwanej (prognozy), jest taki sam. Od razu należy powiedzieć, że w przypadku błędu generalizacji modelu, którego rozkład staramy się oszacować, jest to założenie upraszczające, powoduje ono, że ewentualne informacje uzyskane z rozkładu empirycznego (3.2.11) muszą być traktowane jako przybliżone.

Aby wyjaśnić przyczyny tego stanu rzeczy, przyjrzyjmy się jeszcze raz granicznemu treningowemu błędowi kwadratowemu modelu (3.1.12). Zależność ta została przepisana poniżej jako (3.2.14).

$$Err = \frac{1}{2}\int_{-\infty}^{\infty}(E(y/\mathbf{x}) - f(\mathbf{x},\mathbf{w}))^2 p(\mathbf{x})d\mathbf{x} + \\ + \frac{1}{2}\int_{-\infty}^{\infty}E((y - E(y/\mathbf{x}))^2/\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{3.2.14}$$

Widzimy, że w pierwszej całce w (3.2.14), która w trakcie uczenia decyduje o jakości aproksymacji wartości oczekiwanej rozkładu zmiennej wyjściowej, kwadrat odchylenia ważony jest przez gęstość prawdopodobieństwa wystąpienia danego wzorca wejściowego $p(\mathbf{x})$. W regionach przestrzeni wejść, w których gęstość prawdopodobieństwa $p(\mathbf{x})$ wystąpienia danego wzorca $\mathbf{x}$ jest wysoka, błąd modelu ma więc dużo większy udział w całości średniego błędu kwadratowego niż w regionach, w których gęstość $p(\mathbf{x})$ jest niska, co powoduje powstawanie różnic w jakości aproksymacji (Bishop 1995).

Efekt ten potęguje się w sytuacji, gdy mówimy o konkretnym, skończonym zbiorze danych treningowych, który powinien być dobrą, reprezentatywną próbą wybraną z populacji ogólnej, ale oczywiście nigdy nie będzie próbą idealną. Jeśli pewne regiony przestrzeni wejściowej są słabiej reprezentowane w zbiorze treningowym niż inne, to wyjście modelu dla wzorców wejściowych z tych regionów będzie bardziej odbiegało od faktycznej wartości oczekiwanej prognozy oraz będzie bardziej wrażliwe na wybór innych wzorców z tego regionu, w innych potencjalnych zbiorach treningowych. Innymi słowy, wariancja wyjścia modelu (błąd generalizacji), a przynajmniej jej człon wynikający ze skończonego charakteru zbioru treningowego, a co za tym idzie niepewności parametrów modelu, musi być zależna od konkretnego wzorca wejściowego $\mathbf{x}$.

Musimy więc pamiętać, że rozkład błędów generalizacji nie może być jednakowy dla wszystkich próbek błędu, w związku z tym metodę szacowania rozkładu prognozy opartą na rozkładzie empirycznym błędu należy traktować jako rozwiązanie bardzo przybliżone. Powinniśmy ją zastosować w sytuacji, kiedy nie mamy innego wyjścia, tzn. gdy nie potrafimy określić typu rozkładu

błędu, nie znamy struktury modelu lub nie potrafimy dla danego typu modelu wyznaczyć jego wariancji wyjściowej.

Możemy naturalnie mówić o pewnych rozwiązaniach opartych na rozbiciu przestrzeni wejść modelu na bardziej jednorodne podsegmenty i tworzeniu odrębnych oszacowań empirycznych dla każdego z nich – w ten sposób próbowalibyśmy aproksymować efekt zależności wariancji prognozy od wejścia. Oszacowania takie wymagałyby jednak przeznaczenia na cele testowania modelu tak dużych ilości danych, że nawet w dosyć komfortowych pod tym względem warunkach, z jakimi mamy do czynienia w przypadku krótkoterminowych prognoz zapotrzebowania na energię elektryczną, należy uznać je za mało praktyczne.

W dalszej części prezentowanej pracy skupimy się więc na podejściu parametrycznym do oszacowania warunkowego rozkładu prawdopodobieństwa prognozowanej wielkości dla danego wzorca wejściowego prognozy. Wymaga ono przede wszystkim przyjęcia jakiegoś kształtu rozkładu. Zazwyczaj dobrym pomysłem jest przyjęcie jako modelu błędu rozkładu normalnego. Rozkład normalny Gaussa opisuje zjawiska czysto losowe, ponieważ są one sumą wielu źródeł niepewności. Na podstawie więc centralnego twierdzenia granicznego ich sumaryczny rozkład będzie dążył do rozkładu normalnego. W związku z tym jeżeli mówimy o prognozie średniej (oczekiwanej) wartości pewnej zmiennej, a model jest poprawnie zbudowany i wyjaśnia wszystkie systematyczne czynniki zachowania tej zmiennej, to część niewyjaśniona (czyli błąd) powinna mieć charakter przynajmniej w przybliżeniu losowy, a więc podlegać rozkładowi normalnemu.

Istnieje, rzecz jasna, wiele sytuacji, o których wiemy z teorii, że błąd oszacowania powinien podlegać innemu rozkładowi niż normalny. Jeżeli np. podejmowana decyzja zależy od prognozy wielkości rozproszenia jakiejś zmiennej, to rozkład prognozowanego zjawiska powinien mieć charakter rozkładu F. W przypadku prognozy zjawisk wielokrotnie powtarzanych (niezależnych), należy przyjąć rozkład dwumianowy (lub Poissona).

W punkcie 3.4 zobaczymy jednak, że w przypadku zastosowania sieci neuronowych i neuronowo-rozmytych do krótkoterminowej prognozy zapotrzebowania na energię istnieją pewne podstawy empiryczne do przyjęcia jako modelu błędu rozkładu normalnego. Dlatego w dalszej części naszej monografii będziemy zakładać, że warunkowy rozkład prawdopodobieństwa prognozowanej wielkości y, dla danego wzorca wejściowego $\mathbf{x}$, ma charakter rozkładu normalnego $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$, o wartości oczekiwanej $f(\mathbf{x}, \mathbf{w})$ i odchyleniu standardowym $\sigma_y(\mathbf{x})$:

$$p(y/\mathbf{x}) = \frac{1}{\sigma_y(\mathbf{x})\sqrt{2\pi}} \exp\left(-\frac{(y - f(\mathbf{x}, \mathbf{w}))^2}{2\sigma_y^2(\mathbf{x})}\right) \tag{3.2.15}$$

Wartość oczekiwana rozkładu $f(\mathbf{x}, \mathbf{w})$ równa jest wyjściu modelu (prognozie), zaś wariancja rozkładu $\sigma^2_y(\mathbf{x})$ – wariancji wyjściowej modelu określanej przez komponenty składowe zdefiniowane w zależności (3.2.1) lub w przypadku występowania niepewności wejść – w (3.2.2).

3.2.4. Określanie rozkładu warunkowego prognozy dla modeli regresji liniowej

Zanim przejdziemy do charakterystyki metod szacowania poszczególnych elementów składowych wariancji wyjściowej modelu, w przypadku sieci neuronowych i neuronowo-rozmytych, w kolejnym podpunkcie przyjrzymy się jeszcze metodom wyznaczania rozkładu warunkowego prognozy dla modeli liniowych.

Istnieje kilka powodów, dla których poświęcimy nieco uwagi tej klasie modeli, pomimo że nie stanowi ona głównego tematu naszej pracy. Po pierwsze, modele liniowe posłużą nam jako stosunkowo prosty przykład ilustracyjny wnioskowania na temat parametrów rozkładu warunkowego prognozowanej wielkości, co pozwoli lepiej zrozumieć kilka koncepcji prezentowanych w poprzednich punktach. Po drugie, pamiętać należy, że w niektórych prezentowanych w rozdziale 2 podejściach do krótkoterminowego prognozowania zapotrzebowania na energię i moc wykorzystuje się również modele liniowe. Po trzecie w końcu, i chyba najważniejsze, przedstawione tutaj pewne rozważania dotyczące regresji liniowej uogólnimy i wykorzystamy w niektórych metodach szacowania odchylenia standardowego prognozy dla systemów neuronowych i neuronowo-rozmytych – zostaną one przedstawione w następnych punktach bieżącego rozdziału.

Na początek przedstawmy krótko najważniejsze zagadnienia związane ze sformułowaniem zadania regresji liniowej oraz jego rozwiązaniem. Nasz cel polega na znalezieniu współczynników funkcji liniowej b_i, $i = 0, \ldots, n$ modelującej pewną zależność stochastyczną między wyjściową zmienną objaśnianą (nazywaną również często zmienną zależną) y a wejściowymi zmiennymi objaśniającymi (niezależnymi) $x_1, \ldots, x_n$.

$$y = b_0 + b_1 x_1 + \ldots + b_n x_n + \varepsilon = \mathbf{x}^{\mathrm{T}}\mathbf{b} + \varepsilon = y(\mathbf{x}) + \varepsilon \tag{3.2.16}$$

gdzie $\mathbf{b} = (b_0, b_1, \ldots, b_n)^{\mathrm{T}}$, $\mathbf{x} = (1, x_1, \ldots, x_n)^{\mathrm{T}}$, zaś ε jest czynnikiem losowym zależności. Do znalezienia współczynników powyższej funkcji liniowej metodą najmniejszych kwadratów musimy mieć, oczywiście, próbę danych, odpowiednik zbioru treningowego dla sieci neuronowych lub neuronowo-rozmytych, składającą się ze wzorców wejściowych oraz odpowiadających im znanych (treningowych) wartości zmiennej wyjściowej $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$,

$k = 1, \ldots, N$. Dla każdej znajdującej się w próbie obserwacji tworzymy równanie, które postaramy się rozwiązać ze względu na niewiadome b_i:

$$\begin{aligned} y_1 &= b_0 + b_1 x_{11} + \ldots + b_n x_{1n} \\ y_2 &= b_0 + b_1 x_{21} + \ldots + b_n x_{2n} \\ &\ldots \\ y_N &= b_0 + b_1 x_{N1} + \ldots + b_n x_{Nn} \end{aligned} \tag{3.2.17a}$$

albo w notacji macierzowej:

$$\mathbf{y} = \mathbf{Xb} \tag{3.2.17b}$$

gdzie $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$, $\mathbf{X} = [\mathbf{1}\ x_{ki}]$, k = 1, …, N, i = 1, …, n, zaś $\mathbf{1}$ jest N-elementową kolumną jedynek (czyli $\mathbf{X}$ jest macierzą o wymiarach $\mathrm{N} \times (\mathrm{n}+1)$). Macierz $\mathbf{X}$ często określa się macierzą obserwacji, zaś wektor $\mathbf{y}$ wektorem obserwacji.

Oczywiście utworzony układ równań (3.2.17) nie ma rozwiązania. Gdyby miał, to wszystkie równania musiałyby się układać wzdłuż prostej, czyli zależność (3.2.16) musiałaby mieć charakter funkcyjny, a nie stochastyczny. Zauważmy ponadto, że w układzie (3.2.17) zazwyczaj będziemy mieli znacznie więcej równań (tyle ile obserwacji w próbie – N) niż niewiadomych ($n + 1$ parametrów b_i). Jeśli przyjmiemy, że poszczególne zmienne wejściowe nie są liniowo zależne, co jest jednym z założeń procedury regresji, to rzeczywiste więzy potrzebne do obliczenia niewiadomych, dla dokładnej zależności funkcyjnej między y i $\mathbf{x}$, narzucałyby $n + 1$ równań. Pozostałe równania definiują zależności dodatkowe. Ich liczbę, $N - n - 1$, nazywamy liczbą stopni swobody regresji.

Ponieważ układ równań (3.2.17) nie ma rozwiązania, możemy rozwiązać go jedynie w sposób przybliżony, to znaczy znaleźć takie wartości niewiadomych b_i, aby różnica między prawą i lewą stroną każdego z równań była jak najmniejsza. Zdefiniujmy więc różnicę (błąd) między obiema stronami równań:

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb} \tag{3.2.18}$$

gdzie $\mathbf{e} = (e_1, \ldots, e_N)^{\mathrm{T}}$ jest wektorem odchyleń między wartościami obu stron każdego z równań, nazywanych resztami (albo residuami). Błąd kwadratowy oszacowania zależności (3.2.16), dla danej próby, możemy więc zdefiniować następująco:

$$E = \frac{1}{2}\sum_{k=1}^{N}(y_k - y(\mathbf{x}_k))^2 = \frac{1}{2}\sum_{k=1}^{N}(y_k - \mathbf{x}_k^T\mathbf{b})^2 = \frac{1}{2}\sum_{k=1}^{N}e_k^2 = \frac{1}{2}\mathbf{e}^{\mathrm{T}}\mathbf{e} \tag{3.2.19}$$

Zauważmy przy tym, że:

$$\mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{Xb} - \mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{Xb} = \\ = \mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{Xb} \tag{3.2.20}$$

ponieważ $\mathbf{b}^T\mathbf{X}^T\mathbf{y}$ jest liczbą (skalarem), a więc $\mathbf{b}^T\mathbf{X}^T\mathbf{y} = (\mathbf{b}^T\mathbf{X}^T\mathbf{y})^T = \mathbf{y}^T\mathbf{Xb}$.

Widzimy więc, że błąd (3.2.19) względem parametrów **b** jest funkcją kwadratową. Macierz postaci $\mathbf{X}^T\mathbf{X}$ jest dla dowolnej macierzy **X** dodatnio określona (i dodatkowo symetryczna), a więc błąd (3.2.19) ma dokładnie jedno minimum, które można znaleźć, przyrównując pochodną błędu do 0:

$$-\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{Xb} = 0 \tag{3.2.21a}$$

albo alternatywnie:

$$\mathbf{X}^T\mathbf{Xb} = \mathbf{X}^T\mathbf{y} \tag{3.2.21b}$$

Układ (3.2.21) to układ liniowy nazywany układem równań normalnych. Jeżeli macierz $\mathbf{X}^T\mathbf{X}$ jest nieosobliwa, to układ ten ma rozwiązanie wyznaczające oszacowania wartości parametrów **b** liniowej zależności (3.2.16), minimalizujące błąd kwadratowy (3.2.19), dane wzorem:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.2.22}$$

W praktyce do wyznaczenia parametrów **b** rzadko stosuje się, rzecz jasna, bezpośrednio wzór (3.2.22). Zazwyczaj rozwiązuje się układ równań normalnych numerycznie (3.2.21), stosując szybką metodę rozkładu Cholesky'ego macierzy $\mathbf{X}^T\mathbf{X}$ (która, jak wspomnieliśmy, jest dodatnio określona i symetryczna). Często rezygnuje się w ogóle z generowania równań normalnych, rozwiązując bezpośrednio układ (3.2.17) (w sposób przybliżony, w sensie najmniejszych kwadratów) wolniejszą, ale znacznie bardziej odporną metodą rozkładu na wartości osobliwe (SVD) (Press, Teukolsky, Vetterling, Flannery 1992). Pamiętać jednak należy, że są to tylko różne metody obliczania tej samej wartości **b**, którą mamy w przypadku (3.2.22).

Na tym kończymy nasze wprowadzenie do zagadnień związanych z rozwiązywaniem zadań regresji liniowej. Miało ono naturalnie charakter bardzo skrótowy; zaprezentowaliśmy w nim jedynie główną ideę sposobu otrzymywania oszacowań parametrów. Co do szczegółów oraz wielu innych zagadnień związanych z diagnostyką, analizą i badaniem uzyskanego modelu, odsyłamy Czytelnika do pozycji poświęconych statystyce i analizie danych, takich jak: Draper, Smith 1973; Brandt 1998.

Przejdźmy więc do właściwego tematu naszych zainteresowań, czyli do kwestii szacowania warunkowego rozkładu prawdopodobieństwa wyjścia modelu dla danego wejścia. Pamiętać należy, że procedura stosowana do wyznaczania parametrów modeli liniowych opiera się na metodzie najmniejszych kwadratów, a zatem większość wniosków z wykonanej w punkcie 3.1 analizy błędu kwadratowego, podsumowanych w punkcie 3.2.1, pozostaje w mocy również i w tym przypadku. Jeżeli więc otrzymany model jest poprawnie zbudowany, nieobciążony, to wartość oczekiwana rozkładu prognozowanej zmiennej dana jest przez wyjście modelu (prognozę). Wariancja tego rozkładu składa się z komponentów związanych z wariancją czynnika losowego oraz wariancją modelu wynikającą z parametrów (zależność (3.2.1)). W bieżącym punkcie pominiemy kwestie związane z uwzględnianiem ewentualnej niepewności wejść.

Wnioskowanie o rozkładzie prognozy w przypadku modeli regresji liniowej prowadzi się przy pewnych podstawowych założeniach dotyczących rozkładu błędów resztowych modelu. Zakłada się mianowicie, że błędy te są niezależne oraz stanowią realizacje tego samego błędu losowego ε, o rozkładzie normalnym $N(0, \sigma_\varepsilon)$, czyli o wartości oczekiwanej 0 i stałym odchyleniu standardowym σ_ε. Zauważmy, że w konsekwencji przyjęcia powyższego założenia powodowany przez niego rozkład prawdopodobieństwa pomierzonych wartości y_i, dla poszczególnych $\mathbf{x}_i$, $i = 1, \ldots, N$, w układzie (3.2.17) będzie również miał charakter rozkładu normalnego.

Jeżeli więc spojrzymy na sposób wyznaczania parametrów funkcji liniowej (3.2.22), to widzimy, że obliczane są one za pomocą transformacji liniowej zmiennej losowej **y**, o rozkładzie normalnym. Wyznaczone za pomocą tej zależności parametry **b** obarczone są więc wynikającą z tego faktu niepewnością, opisywaną wielowymiarowym normalnym rozkładem prawdopodobieństwa $N(\mathbf{b}, \mathbf{C_b})$:

$$p(\boldsymbol{\beta}) = k \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})^{\mathrm{T}} \mathbf{C}_{\mathbf{b}}^{-1} (\boldsymbol{\beta} - \mathbf{b})\right) \tag{3.2.23}$$

gdzie wyznaczone metodą najmniejszych kwadratów parametry **b** są wartościami oczekiwanymi rozkładu, $\mathbf{C_b}$ macierzą kowariancji oszacowanych parametrów (musimy ją znaleźć), zaś k współczynnikiem normalizacyjnym, którego dokładna postać nie będzie nas interesowała. Analizując dalej wpływ tej niepewności na wyjście modelu $y(\mathbf{x})$, dla danego wzorca wejściowego **x**, tj. $y(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{b}$, ponownie zauważmy, że dla ustalonego wejścia jest ono funkcją liniową parametrów **b**. Innymi słowy, warunkowy rozkład prognozy modelu spowodowany niepewnością parametrów będzie miał (przy przyjętych założeniach) charakter rozkładu normalnego $N(\mathbf{x}^{\mathrm{T}}\mathbf{b}, \sigma_b(\mathbf{x}))$, którego wartość oczekiwana równa jest wyjściu modelu (co wynika z analizy błędu kwadratowego i wynika-

jącej z niej zależności (3.1.13)), a odchylenie standardowe $\sigma_b(\mathbf{x})$ będziemy musieli oszacować.

Wróćmy jeszcze na chwilę do założonego rozkładu reszt $N(0, \sigma_\varepsilon)$. Zauważmy, że reszty, jako błędy popełnione przez model na próbie, nie zawierają w sobie błędu generalizacji, ponieważ mierzone są dla tych samych wzorców danych, które wykorzystane zostały do dopasowania modelu. Wobec tego, zakładając, że sam model jest nieobciążony, czyli mówiąc w skrócie, jego wyjście stanowi dobrą aproksymację wartości oczekiwanej prognozy, sam rozkład reszt modelu jako taki stanowić będzie oszacowanie wyłącznie rozkładu czynnika losowego (szumu losowego) zależności między zmiennymi objaśniającymi a objaśnianą, nieuwzględniające niepewności samego modelu.

Podsumowując, przyjęcie założenia o normalnym rozkładzie błędów resztowych modelu pociąga za sobą normalny charakter dwóch podstawowych niezależnych źródeł niepewności prognozy wykonywanej za pomocą liniowego modelu regresyjnego: niepewności wyjścia modelu wynikającej ze skończonego charakteru próby i powstającej w efekcie niepewności parametrów modelu, opisanej rozkładem $N(\mathbf{x}^{\mathrm{T}}\mathbf{b}, \sigma_b(\mathbf{x}))$, oraz czynnika losowego, opisanego rozkładem $N(0, \sigma_\varepsilon)$. W związku z tym warunkowy rozkład prognozy dla danego wejścia modelu będzie miał również charakter rozkładu normalnego $N(\mathbf{x}^{\mathrm{T}}\mathbf{b}, \sigma_y(\mathbf{x}))$:

$$p(y/\mathbf{x}) = \frac{1}{\sigma_y(\mathbf{x})\sqrt{2\pi}} \exp\left(-\frac{(y-\mathbf{x}^T\mathbf{b})^2}{2\sigma_y^2(\mathbf{x})}\right) \tag{3.2.24}$$

gdzie $\sigma^2_y(\mathbf{x})$ jest wariancją wyjściową modelu określoną przez:

$$\sigma_y^2(\mathbf{x}) = \sigma_b^2(\mathbf{x}) + \sigma_\varepsilon^2 \tag{3.2.25}$$

Aby więc uzyskać oszacowanie rozkładu prawdopodobieństwa prognozy modelu (3.2.24), musimy wyznaczyć wartości wariancji $\sigma^2_b(\mathbf{x})$ i σ^2_ε. Rozpocznijmy od pierwszej z nich, czyli od wariancji wynikającej z niepewności parametrów. Wyznaczymy ją dwukrokowo, znajdując macierz kowariancji $\mathbf{C}_b$ parametrów modelu, a następnie wariancję wyjściową $\sigma^2_b(\mathbf{x})$ wywoływaną przez tę kowariancję. W obydwu przypadkach potrzebne oszacowania uzyskuje się praktycznie natychmiast na podstawie jednego z podstawowych praw wykorzystywanych w analizie danych i miernictwie, mianowicie prawa propagacji błędów. Ponieważ jego wyprowadzenie jest niezbyt długie i nieskomplikowane, a w przyszłości będziemy z niego jeszcze korzystać, więc przedstawimy je pokrótce poniżej.

Przyjmijmy, że w ogólnym przypadku mamy pewne zmienne losowe $\mathbf{z} = (z_1, \ldots, z_k)$, które są wynikiem transformacji liniowej zmiennych $\mathbf{t} = (t_1, \ldots, t_r)$, tzn.:

$$\mathbf{z} = \mathbf{A}\mathbf{t} + \mathbf{b} \tag{3.2.26}$$

Wówczas z podstawowych właściwości wartości oczekiwanej rozkładu prawdopodobieństwa mamy:

$$E(\mathbf{z}) = \mathbf{A}E(\mathbf{t}) + \mathbf{b} \tag{3.2.27}$$

Natomiast macierz kowariancji transformowanych zmiennych **z** możemy wyznaczyć w następujący sposób:

$$\begin{aligned}
\mathbf{C}_z &= E\left((\mathbf{z} - E(\mathbf{z}))(\mathbf{z} - E(\mathbf{z}))^T\right) = \\
&= E\left((\mathbf{A}\mathbf{t} + \mathbf{b} - \mathbf{A}E(\mathbf{t}) - \mathbf{b})(\mathbf{A}\mathbf{t} + \mathbf{b} - \mathbf{A}E(\mathbf{t}) - \mathbf{b})^T\right) = \\
&= E\left((\mathbf{A}\mathbf{t} - \mathbf{A}E(\mathbf{t}))(\mathbf{A}\mathbf{t} - \mathbf{A}E(\mathbf{t}))^T\right) = \\
&= E\left(\mathbf{A}(\mathbf{t} - E(\mathbf{t}))(\mathbf{t} - E(\mathbf{t}))^T\mathbf{A}^T\right) = \\
&= \mathbf{A}E\left((\mathbf{t} - E(\mathbf{t}))(\mathbf{t} - E(\mathbf{t}))^T\right)\mathbf{A}^T \\
&= \mathbf{A}\mathbf{C}_t\mathbf{A}^T
\end{aligned} \tag{3.2.28}$$

Zależność (3.2.28) daje nam tzw. prawo propagacji błędów, które pokazuje, jak zmieniają się błędy (wariancje) zmiennych wejściowych **t** po ich transformacji na zmienne wyjściowe **z** za pomocą odwzorowania liniowego o macierzy **A**:

$$\mathbf{C}_z = \mathbf{A}\mathbf{C}_t\mathbf{A}^T \tag{3.2.29}$$

gdzie $\mathbf{C}_z$, $\mathbf{C}_t$ są macierzami kowariancji, odpowiednio zmiennych **z** i **t**.

Przypomnijmy teraz, że optymalne wartości parametrów **b** wyznaczane są za pomocą zależności (3.2.22), na podstawie **y**, wektora obserwowanych w próbie wartości zmiennej objaśnianej y. Zauważmy, że zależność ta ma charakter transformacji liniowej o macierzy odwzorowania $\mathrm{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. Ponieważ założyliśmy, że rozkład błędów resztowych e_i, i = 1, …, N ma stałą wariancję σ^2_{ε}, to podobnie będzie również w przypadku poszczególnych y_i. Ponadto zakładaliśmy, że poszczególne reszty są niezależne, niezależne więc będą także y_i, a więc kowariancja $\mathrm{cov}(y_i, y_j) = 0$ dla $i \neq j$. Macierz kowariancji zmiennej wielowymiarowej **y**, złożonej ze wszystkich obserwowanych wartości y_i, i = 1, …, N, będzie więc miała na głównej przekątnej wartości σ^2_{ε}, a poza nią zera:

$$\mathbf{C}_\mathbf{y} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & \ldots & 0 \\ 0 & \sigma_\varepsilon^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_\varepsilon^2 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I} \tag{3.2.30}$$

gdzie **I** jest macierzą jednostkową, złożoną z jedynek na głównej przekątnej i zer poza nią.

Macierz $\mathbf{C}_b$ kowariancji parametrów naszego modelu **b**, uzyskanych dla danego zbioru danych (próby), możemy więc wyznaczyć, stosując prawo propagacji błędów do macierzy kowariancji $\mathbf{C}_\mathbf{y}$:

$$\begin{aligned} \mathbf{C}_b &= \mathbf{A}\mathbf{C}_\mathbf{y}\mathbf{A}^\mathrm{T} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\,\mathbf{C}_\mathbf{y}\,((\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T})^\mathrm{T} = \\ &= (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\,\mathbf{I}\sigma_\varepsilon^2\,\mathbf{X}((\mathbf{X}^\mathrm{T}\mathbf{X})^{-1})^\mathrm{T} = \\ &= \sigma_\varepsilon^2(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\,\mathbf{I}\,\mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1} = \\ &= \sigma_\varepsilon^2\,(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1} = \sigma_\varepsilon^2\,(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1} \end{aligned} \tag{3.2.31}$$

Wykorzystaliśmy tutaj fakt, że macierz $\mathbf{X}^\mathrm{T}\mathbf{X}$ jest symetryczna, w związku z tym symetryczna jest również macierz $(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$. A zatem $((\mathbf{X}^\mathrm{T}\mathbf{X})^{-1})^\mathrm{T} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$.

Dla danego wejścia modelu **x** jego wyjście $y(\mathbf{x})$ również otrzymywane jest w wyniku przekształcenia liniowego parametrów $y(\mathbf{x}) = \mathbf{x}^\mathrm{T}\mathbf{b}$. Zmienna y ma charakter liczbowy, a więc wyjściowa macierz kowariancji redukuje się wyłącznie do jednego elementu, tj. wariancji wyjściowej modelu spowodowanej niepewnością parametrów: $\sigma_b^2(\mathbf{x})$. Wyznaczyć możemy ją ponownie na mocy prawa propagacji błędów, przyjmując jako macierz $\mathbf{A} = \mathbf{x}^\mathrm{T}$. I tak niemal natychmiast otrzymujemy:

$$\sigma_b^2(\mathbf{x}) = \mathbf{A}\mathbf{C}_b\mathbf{A}^\mathrm{T} = \mathbf{x}^\mathrm{T}\mathbf{C}_b\mathbf{x} = \mathbf{x}^\mathrm{T}\sigma_\varepsilon^2(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{x} = \sigma_\varepsilon^2\,\mathbf{x}^\mathrm{T}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{x} \tag{3.2.32}$$

Wracając do zależności (3.2.25), stwierdzamy, że łączna wariancja wyjścia modelu $y(\mathbf{x})$ dla danego wzorca wejściowego, uwzględniająca zarówno wariancję powodowaną przez niepewność parametrów oraz element czynnika losowego, wynosi:

$$\sigma_y^2(\mathbf{x}) = \sigma_\varepsilon^2 + \sigma_b^2(\mathbf{x}) = \sigma_\varepsilon^2 + \sigma_\varepsilon^2\,\mathbf{x}^\mathrm{T}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{x} = \sigma_\varepsilon^2\,(1 + \mathbf{x}^\mathrm{T}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{x}) \tag{3.2.33}$$

Ostatecznie więc możemy przyjąć, że warunkowy rozkład prognozowanej wielkości, dla danego wejścia, ma charakter rozkładu normalnego:

$$p(y/\mathbf{x}) = N\left(\mathbf{x}^T\mathbf{b},\ \ \sigma_\varepsilon\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}\right) \tag{3.2.34}$$

Zwróćmy jeszcze uwagę, że aby w praktyce korzystać z rozkładu (3.2.34), musimy znaleźć sposób oszacowania pojawiającego się w nim odchylenia standardowego (lub wariancji) rozkładu prawdopodobieństwa czynnika losowego σ_ε. Przypomnijmy, że realizacją serii danych z tego rozkładu jest zbiór błędów resztowych modelu w próbie. W związku z tym oszacowanie σ_ε otrzymuje się za pomocą standardowego estymatora odchylenia standardowego błędu, pierwiastka średniego kwadratu reszt modelu (błędu standardowego w próbie) S_{N-n-1}:

$$S_{N-n-1} = \sqrt{\frac{1}{N-n-1}\sum_{i=1}^{N} e_i^2} = \sqrt{\frac{1}{N-n-1}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T b)^2} \tag{3.2.35}$$

gdzie N jest licznością próby, n liczbą zmiennych objaśniających (n + 1 liczbą parametrów modelu), tak więc $N - n - 1$ jest liczbą stopni swobody.

Korzystając z otrzymanego oszacowania (3.2.33) rozkładu prawdopodobieństwa warunkowego rozkładu prognozowanej wielkości, możemy również łatwo wyznaczyć dla danej prognozy przedział jej wartości, przy przyjętym poziomie prawdopodobieństwa α. Na podstawie (3.2.5) (punkt 3.2.2), dla konkretnego wejścia prognozy $\mathbf{x}$, krańce przedziału, w którym prognozowana zmienna znajdować będzie się z prawdopodobieństwem α, odpowiednio są równe:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= Q_{N\left(\mathbf{x}^T\mathbf{b},\sigma_y(\mathbf{x})\right)}((1-\alpha)/2) = Q_{N\left(\mathbf{x}^T\mathbf{b},\ \sigma_\varepsilon\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}\right)}((1-\alpha)/2) \\ g_{y/\mathbf{x}}(\alpha) &= Q_{N\left(\mathbf{x}^T\mathbf{b},\sigma_y(\mathbf{x})\right)}((1+\alpha)/2) = Q_{N\left(\mathbf{x}^T\mathbf{b},\ \sigma_\varepsilon\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}\right)}((1+\alpha)/2) \end{aligned} \tag{3.2.36}$$

gdzie $Q(\alpha)$ jest kwantylem warunkowego rozkładu prawdopodobieństwa prognozowanej wielkości, dla prawdopodobieństwa α.

Oczywiście w przypadku obliczeń ręcznych i z wykorzystaniem tablic statystycznych przedział prognozy możemy wyznaczyć, stosując wartości kwantyli rozkładu normalnego standardowego $N(0, 1)$:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= \mathbf{x}^T\mathbf{b} - Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_\varepsilon\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}} \\ g_{y/\mathbf{x}}(\alpha) &= \mathbf{x}^T\mathbf{b} + Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_\varepsilon\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}} \end{aligned} \tag{3.2.37}$$

Zauważmy, że w (3.2.37) przy wyznaczaniu dolnego krańca przedziału prognozy korzystamy z symetrii rozkładu normalnego standardowego, wokół wartości oczekiwanej 0. W związku z tym $Q_{N(0,1)}((1-\alpha)/2) = -Q_{N(0,1)}((1+\alpha)/2)$.

Jeżeli jako oszacowanie odchylenia standardowego szumu losowego σ_ε wykorzystywane jest jego oszacowanie na podstawie błędów modelu z próby (3.2.35), to do wyznaczenia przedziałów prognozy powinniśmy skorzystać z kwantyli rozkładu t-Studenta, zamiast z rozkładu normalnego:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= \mathbf{x}^T\mathbf{b} - t_{N-n-1}((1+\alpha)/2)\cdot S_{N-n-1}\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}} \\ g_{y/\mathbf{x}}(\alpha) &= \mathbf{x}^T\mathbf{b} + t_{N-n-1}((1+\alpha)/2)\cdot S_{N-n-1}\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}} \end{aligned} \tag{3.2.38}$$

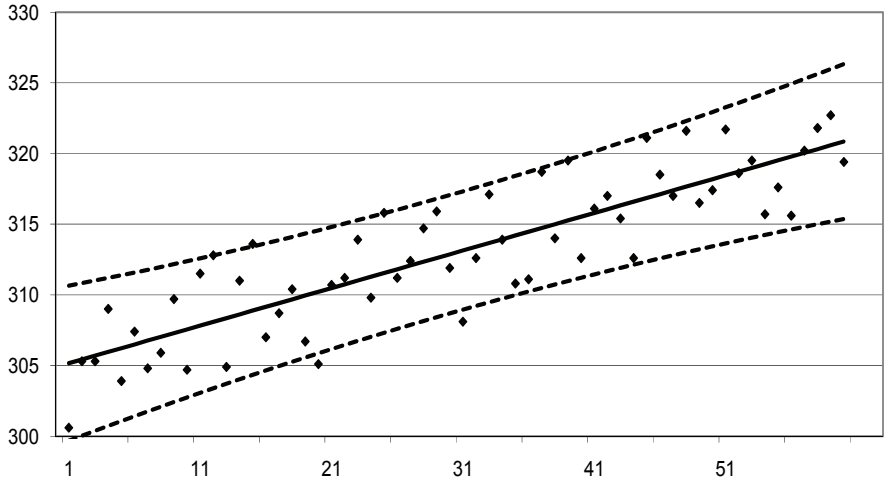
gdzie $t_{N-n-1}((1+\alpha)/2)$ jest jednostronnym kwantylem rozkładu t-Studenta, dla prawdopodobieństwa $(1+\alpha)/2$ i $N-n-1$ stopni swobody. Zamiana rozkładu normalnego na rozkład t-Studenta jest naturalnie istotna w sytuacji niewielkiej liczby stopni swobody, to znaczy kiedy rozmiar próby niewiele przekracza liczbę parametrów. W przypadku prognozy krótkoterminowego zapotrzebowania na energię elektryczną, gdzie modele tworzone są zazwyczaj na próbach (zbiorach treningowych) złożonych z kilkuset wzorców, rozkład Studenta w zasadzie pokrywa się z normalnym.

Uważny Czytelnik może zadać sobie pewne pytanie. Czy przyjęcie stałego rozkładu $N(0, \sigma_\varepsilon)$ wszystkich błędów resztowych modelu jest założeniem poprawnym? Przypomnijmy, że kiedy mówiliśmy w punkcie 3.2.3 o rozkładzie empirycznym prognozy, głównym problemem występującym w tym rozwiązaniu było właśnie założenie o jednakowym rozkładzie wszystkich próbek błędu modelu. W efekcie otrzymany rozkład empiryczny prognozowanej wielkości również był stały i taki sam dla każdej prognozy. Dyskutowaliśmy dalej, że może być to poważne uproszczenie, ponieważ rozkład ten powinien mieć charakter warunkowy, zależny od wykorzystanego wzorca wejściowego prognozy, a przede wszystkim jego wariancja powinna być zależna od danego wejścia $\mathbf{x}$.

Zwróćmy jednak uwagę, że dyskusja w punkcie 3.2.3 dotyczyła rozkładu błędów generalizacji, natomiast założenie poczynione w bieżącym rozdziale związane jest z rozkładem błędów w próbie, a więc dotyczy wyłącznie niepewności czynnika losowego. Otrzymane dalej we wzorze (3.2.33) oszacowanie wariancji łącznego rozkładu wyjścia modelu $\sigma^2_y(\mathbf{x})$, uwzględniającej zarówno niepewność czynnika losowego, jak i niepewność samego modelu, jest już zależne od wzorca wejściowego $\mathbf{x}$. W konsekwencji warunkowy rozkład prognozowanej wielkości dla każdego $\mathbf{x}$ ma, co prawda, charakter rozkładu normalnego, ale o zmiennej wariancji.

Czy na pewno wariancja rozkładu wyznaczona przy użyciu (3.2.33) zmienia się w zależności od $\mathbf{x}$? Pokażemy, że tak jest na przykładzie. Na rysunku 3.2.1 przedstawiona została (ciągła linia) prosta $y = b_1x + b_0$, dopasowana metodą

najmniejszych kwadratów do zbioru punktów, oraz krańce przedziału prognozy dla uzyskanego modelu liniowego, wyznaczone na podstawie wzoru (3.2.38), dla prawdopodobieństwa 95% (linie przerywane). Jeżeli przyjrzymy się szerokości wykreślonych przedziałów prognozy, zauważymy bez trudu, że są one nieco szersze na krańcach zbioru wartości zmiennej x niż w pobliżu jego środka.



Rysunek 3.2.1. Przykładowa prosta dopasowana do zbioru punktów i przedział prognozy dla prawdopodobieństwa $\alpha = 95\%$
Źródło: opracowanie własne

Dokładnie rzecz ujmując, krańce przedziałów prognozy (determinowane przez wariancję rozkładu wyjściowego modelu) w funkcji zmiennej x mają kształt hiperboliczny. Najmniejszą szerokość (najniższa wariancja rozkładu warunkowego y / x) osiągają dla średniej $x = \overline{x}$; im bardziej oddalamy się od punktu środkowego danych (w dowolną stronę), tym bardziej rośnie wariancja rozkładu, a co za tym idzie – rosną przedziały prognozy. Innymi słowy, oddalając się od środka danych (średniej argumentów), spodziewamy się popełnić większy błąd w oszacowaniu wartości oczekiwanej $E(y / x)$, przy użyciu prostej dopasowanej do tych danych. Wychodząc zupełnie poza przedział obserwacji, dla wartości x poza zakresem danych, możemy się spodziewać dalszego pogarszania się dokładności prognozy (Draper, Smith 1973).

Sytuację tę łatwo zinterpretować i uzasadnić. Ma ona zresztą również dobrą podbudowę intuicyjną. Najlepszych wyników prognozy spodziewamy się w „środku" danych, gdzie położenie prostej (wartości parametrów) zdetermino-

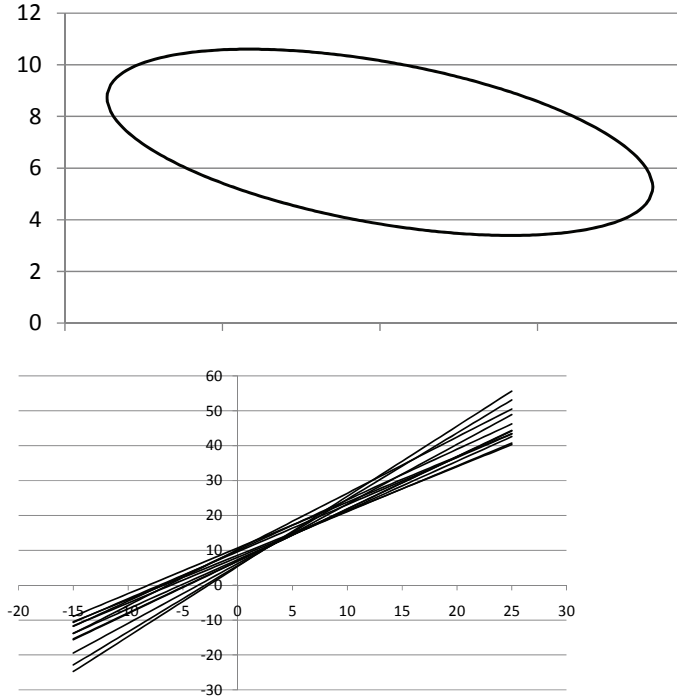
wane jest przez dane ze środka ich zakresu oraz mniej więcej w jednakowym stopniu z obu jego krańców. Położenie prostej na którymś z krańców zbioru danych determinowane jest silnie przez dane z tego krańca, ze środka już słabiej. Ewentualna zmiana parametrów, wynikająca z ich niepewności, musi więc być w pewien sposób zsynchronizowana, tak by potencjalne proste definiowane przez różne modele skupiały się, czy nawet przecinały, w okolicach centrum danych, tworząc w nim coś w rodzaju „środka obrotu", wywoływanego przez zmiany w wartości współczynnika kierunkowego prostej regresji. W takiej sytuacji nawet stosunkowo małe zmiany współczynnika kierunkowego, powodujące niewielkie przesunięcia względem osi y w centrum danych, dają w efekcie dużo większe przemieszczenia na krańcach ich zakresu.

Opisane zjawisko można oczywiście wyjaśnić teoretycznie. Przypomnijmy, że rozkład prawdopodobieństwa możliwych parametrów modelu $\boldsymbol{\beta}$, wokół optymalnych parametrów $\mathbf{b}$, oszacowanych metodą najmniejszych kwadratów, ma charakter wielowymiarowego rozkładu normalnego $N(\mathbf{b}, \mathbf{C_b})$, określonego wzorem (3.2.23). Macierz kowariancji parametrów możemy wyznaczyć za pomocą wzoru (3.2.31). Jeżeli chcemy znaleźć wartości parametrów odpowiadające danemu poziomowi prawdopodobieństwa, musimy odszukać kontur w przestrzeni parametrów, dla którego wykładnik funkcji gęstości rozkładu (3.2.23) równy będzie stałej wartości c odpowiadającej temu prawdopodobieństwu:

$$(\boldsymbol{\beta} - \mathbf{b})^{\mathrm{T}} \mathbf{C}_{\mathbf{b}}^{-1} (\boldsymbol{\beta} - \mathbf{b}) = c \tag{3.2.39}$$

Na płaszczyźnie dla dwóch parametrów zależność (3.2.39) stanowi równanie elipsy o środku w punkcie $\mathbf{b} = (b_0, b_1)$, nazywanej elipsą ufności – zależność ta wyznaczana jest przez poziomy przekrój (poziomicę) dwuwymiarowego rozkładu normalnego Gaussa. Linie stałego prawdopodobieństwa dla par parametrów są więc wzajemnie koncentrycznymi elipsami, których rozmiar rośnie w miarę spadku wartości prawdopodobieństwa. W przypadku modeli o większej liczbie parametrów elipsy zostaną zastąpione oczywiście przez wielowymiarowe hiperpowierzchnie o eliptycznym kształcie.

Na rysunku 3.2.2 przedstawiona została przykładowa elipsa ufności dla dwóch parametrów oraz grupa kilku wybranych linii prostych, których parametry zmieniają się zgodnie z zależnością odpowiadającą konturowi tej elipsy. Wystąpienie każdej z tych prostych byłoby jednakowo prawdopodobne dla danego zbioru danych (próby), na podstawie którego otrzymano elipsę ufności. Jak widzimy w dolnej części rysunku, otrzymane na podstawie elipsy proste wyraźnie położone są bliżej siebie w okolicy środka przedziału argumentu, znacznie bardziej rozpraszając się w pobliżu jego krańców.



Rysunek 3.2.2. Przykład elipsoidy kowariancji parametrów funkcji liniowej i rodziny prostych, których parametry zmieniają się ze stałym prawdopodobieństwem na konturze tej elipsy
Źródło: opracowanie własne

Zastanówmy się jeszcze nad związkiem między elipsoidą ufności (czyli rozkładem parametrów modelu) a wartościami funkcji błędu kwadratowego E. Przypomnijmy, że, jak to wcześniej pokazaliśmy w zależności (3.2.20), błąd kwadratowy (3.2.19) dla danego zestawu parametrów modelu β możemy zapisać:

$$E(\boldsymbol{\beta}) = \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \tfrac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}) \qquad (3.2.40)$$

Natomiast minimalna wartość tego błędu, dla optymalnych parametrów **b** wyznaczonych z układu równań normalnych (3.2.21), wynosi:

$$\begin{aligned} E(\mathbf{b}) &= \tfrac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}) = \\ &= \tfrac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}) = \tfrac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}) \end{aligned} \qquad (3.2.41)$$

ponieważ **b** spełnia układ równań normalnych (3.2.21), czyli $\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = 0$.

Jeżeli teraz obliczymy różnicę między wartością błędu dla dowolnego zestawu parametrów $\boldsymbol{\beta}$ a jego minimum dla parametrów optymalnych, to otrzymamy:

$$E(\boldsymbol{\beta}) - E(\mathbf{b}) = \tfrac{1}{2}(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) - \tfrac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}) = \\ = \tfrac{1}{2}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{b}^T\mathbf{X}^T\mathbf{y}) \tag{3.2.42}$$

Jeszcze raz wykorzystajmy fakt, że na podstawie układu równań normalnych (3.2.21) dla optymalnych parametrów $\mathbf{b}$ zachodzi zależność $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{b}$, więc różnica błędów (3.2.42) wynosi:

$$E(\boldsymbol{\beta}) - E(\mathbf{b}) = \tfrac{1}{2}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{b}^T\mathbf{X}^T\mathbf{y}) = \\ = \tfrac{1}{2}(-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\mathbf{b} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}) = \\ = \tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \tag{3.2.43}$$

Przypomnijmy sobie, że we wzorze (3.2.31) wyznaczyliśmy macierz kowariancji parametrów liniowego modelu regresji $\mathbf{C}_\mathbf{b} = \sigma^2_\varepsilon(\mathbf{X}^T\mathbf{X})^{-1}$. Wobec tego ostatecznie różnicę między wartością błędu kwadratowego dla dowolnego zestawu parametrów $\boldsymbol{\beta}$ a jego minimum dla parametrów optymalnych $\mathbf{b}$ możemy zapisać:

$$E(\boldsymbol{\beta}) - E(\mathbf{b}) = \frac{1}{2\sigma^2_\varepsilon}(\boldsymbol{\beta} - \mathbf{b})^T\mathbf{C}_\mathbf{b}^{-1}(\boldsymbol{\beta} - \mathbf{b}) \tag{3.2.44}$$

Zauważmy więc, że zmiana wartości błędu (3.2.44) odpowiada elipsoidzie ufności (3.2.39). Dokładniej rzecz biorąc, można pokazać, że elipsoida ufności odpowiadająca prawdopodobieństwu p jest pewną hiperpowierzchnią w przestrzeni parametrów modelu, na której wartość funkcji błędu kwadratowego wynosi:

$$E(\boldsymbol{\beta}) = E(\mathbf{b}) + \chi^2_p(N - n - 1) \tag{3.2.45}$$

gdzie χ^2_p jest kwantylem rozkładu χ^2 o $N - n - 1$ stopniach swobody (liczba obserwacji w próbie minus liczba parametrów modelu), dla prawdopodobieństwa p (Brandt 1998). Ten istotny z punktu widzenia weryfikacji modelu liniowego regresji fakt interesować nas będzie nieco mniej. W naszym przypadku ważniejsza okaże się zależność (3.2.24) warunkująca zmianę błędu dla różnych wartości parametrów od macierzy kowariancji parametrów. Tę informację będziemy wykorzystywać już w następnym podrozdziale.

Na koniec bieżącego punktu przypomnijmy jeszcze raz, że wszystkie wyniki związane z analizą rozkładu warunkowego prawdopodobieństwa prognozo-

wanej wartości uzyskane były przy dwóch podstawowych założeniach dotyczących błędu resztowego modelu. Przede wszystkim zakładaliśmy, że poszczególne błędy mają charakter niezależny od siebie, oraz że rozkład reszt, a, co za tym idzie, błędu losowego jest rozkładem normalnym $N(0, \sigma_\varepsilon)$. Niezależność błędów oraz założenie o zerowej wartości oczekiwanej błędu losowego $E(\varepsilon) = 0$ zasadniczo wchodzą w skład standardowego zestawu założeń szacowania parametrów modelu liniowego metodą najmniejszych kwadratów, których tutaj szeroko nie omawiamy, odsyłając Czytelnika do odpowiedniej literatury przedmiotu (np. Draper, Smith 1973; Brandt 1998). Założenie odnośnie do normalnego charakteru rozkładu prawdopodobieństwa błędów jest jednak założeniem dodatkowym, niezbędnym wyłącznie do wnioskowania na temat rozkładu wyjścia modelu dla danego wzorca wejściowego.

Dla poprawnie zbudowanego modelu, w zadaniach związanych z prognozowaniem zapotrzebowania na energię (moc), poczynienie takiego założenia, jak już wspominaliśmy w punkcie 3.2.3, na ogół jest rozsądne. Często rozkład tego typu prognoz – prognoz wartości oczekiwanej jakiejś zmiennej – będzie miał charakter zbliżony do normalnego. Jednakże automatyczne przyjmowanie rozkładu normalnego błędu wyłącznie na podstawie centralnego twierdzenia granicznego może być postępowaniem zbyt optymistycznym. Należy je zweryfikować empirycznie, sprawdzając rozkład reszt naszego modelu prognostycznego za pomocą któregoś ze znanych standardowych testów normalności rozkładu (takich jak test Jarque–Bera czy Kołmogorowa–Smirnowa). Jeżeli odbiega on poważnie od rozkładu normalnego, wyniki wnioskowania odnośnie do rozkładu prawdopodobieństwa prognozy należy traktować z dużą nieufnością.

Możemy mówić o pewnej możliwości złagodzenia założenia o stałej wariancji błędu resztowego (losowego). Jeżeli błąd losowy będzie miał rozkład normalny $N(0, \sigma_\varepsilon(\mathbf{x}))$, o zmiennej wariancji zależnej od wejścia $\mathbf{x}$ (tzw. heteroskedastyczność modelu), to nasze rozważania możemy uznać za w przybliżeniu poprawne. Oczywiście w kilku miejscach założyliśmy stały charakter błędu σ_ε. Przede wszystkim samo dopasowanie prostej regresji powinno odbywać się przy wykorzystaniu tzw. metody ważonych najmniejszych kwadratów, gdzie poszczególne człony błędu w błędzie kwadratowym dzielone są przez wariancję błędu dla danej wartości:

$$E = \frac{1}{2}\sum_{k=1}^{N}\frac{(y_k - \mathbf{x}_k^T\mathbf{b})^2}{\sigma_k^2(\mathbf{x})} \tag{3.2.46}$$

Problem polega na tym, że wówczas odchylenie standardowe rozkładu błędu $\sigma_\varepsilon(\mathbf{x})$ musiałoby być znane z góry, co w zagadnieniach prognostycznych byłoby trudne do spełnienia. Jeżeli jednak model jest heteroskedastyczny,

odchylenie standardowe błędu się zmienia, ale jego rozkład jest normalny $N(0, \sigma_\varepsilon(\mathbf{x}))$, to symetryczny charakter rozkładu powinien zapewnić poprawność dopasowania.

Również nasze rozważania dotyczące rozkładu wyjścia modelu powinny lokalnie, w otoczeniu punktu **x**, pozostawać w przybliżeniu właściwe. Nie możemy wówczas, rzecz jasna, używać jako estymatora odchylenia standardowego błędu stałej wartości S_{N-n-1} określonej wzorem (3.2.35). Musimy dostarczyć nowego oszacowania odchylenia standardowego (lub wariancji) $\sigma_\varepsilon(\mathbf{x})$, zależnego od wejścia modelu **x**. Zagadnienia tego typu będziemy omawiać w punkcie 3.4.

3.3. Wyznaczanie wariancji prognozy wynikającej z niepewności parametrów modelu neuronowego (neuronowo-rozmytego)

3.3.1. Podejścia do szacowania wariancji wyjściowej modelu z parametrów w przypadku nieliniowym

Z analizy wykonanej w początkowych punktach bieżącego rozdziału wynika, że w przypadku modeli dopasowywanych do danych historycznych metodą najmniejszych kwadratów, tak jak to jest w przypadku wykorzystywanych przez nas sieci neuronowych (neuronowo-rozmytych), warunkowy rozkład prognozowanego zapotrzebowania na energię (moc) wyznaczany zostaje przez rozkład prawdopodobieństwa zmian wyjścia modelu prognostycznego, wokół wartości oczekiwanej zapotrzebowania określanej przez obserwowaną wartość prognozy. Zakładając więc normalny charakter rozkładu błędu modelu, możemy przyjąć, że prognozowana wielkość opisana jest rozkładem normalnym prawdopodobieństwa $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$, o wartości oczekiwanej $f(\mathbf{x}, \mathbf{w})$ (wyjście sieci) i odchyleniu standardowym $\sigma_y(\mathbf{x})$. Przypomnijmy jeszcze, że wariancja wyjściowa modelu $\sigma^2_y(\mathbf{x})$ zgodnie ze wzorem (3.2.2) określana jest przez trzy podstawowe elementy (dokładniej mówiąc, ponieważ są to niezależne źródła błędu, stanowi ich sumę): $\sigma^2_{\mathbf{w}}(\mathbf{x})$ wariancję wyjścia modelu prognostycznego wynikającą z niepewności parametrów (wag), $\sigma^2_\varepsilon(\mathbf{x})$ wariancję czynnika losowego (szum losowy) oraz $\sigma^2_{\mathbf{x}}(\mathbf{x})$ wariancję wyjścia modelu spowodowaną niepewnością wejść (jeśli taka niepewność występuje).

W bieżącym podrozdziale zajmiemy się sposobami szacowania pierwszego ze wspomnianych komponentów rozkładu prognozy, tzn. wariancji wyjścia modelu neuronowego (neuronowo-rozmytego) $\sigma^2_{\mathbf{w}}(\mathbf{x})$ wynikającej z niepewności parametrów. Od razu należy tutaj zaznaczyć, że w przypadku nieliniowym, a o takim przecież mówimy, nie mamy do dyspozycji tak dokładnej i precyzyjnej teorii, jaką oferowały nam liniowe modele regresji (patrz punkt 3.2.4).

Metody omawiane w bieżącym podrozdziale polegają więc na różnego rodzaju uproszczeniach, przybliżeniach lub zastosowaniu empirycznych oszacowań na podstawie danych. Efekty naszych badań, zaprezentowane w kolejnych podpunktach, wskazują naturalnie, że otrzymywane za ich pomocą wyniki, aczkolwiek przybliżone, dają przecież w miarę dokładne oszacowania rozkładu prognozy zapotrzebowania na energię. Należy jednak pamiętać, że każda z omawianych metod powinna zostać przed praktycznym zastosowaniem zweryfikowana dla konkretnego przypadku.

W literaturze dotyczącej zagadnienia szacowania warunkowego rozkładu predykcji nieliniowego modelu neuronowego lub neuronowo-rozmytego prezentuje się kilka różnych metod, które można wykorzystać do tego zadania. Stosowane podejścia ogólnie możemy podzielić na trzy następujące grupy:

1. Oszacowania analityczne, takie jak: metoda delta (*delta method*) (Chryssolouris, Lee, Ramsey 1996; Dybowski, Roberts 1999; Penny, Roberts 1997; Tibshirani 1996), estymator kanapkowy (*sandwich estimator*) (Tibshirani 1996). Opierają się na analizie hesjanu (macierzy drugich pochodnych) błędu względem parametrów modelu. Wymagają pewnych przybliżeń związanych z lokalną linearyzacją modelu, co pozwala na zastosowanie zmodyfikowanej teorii dla modeli liniowych.

2. Oszacowania oparte na eksperymentach Monte Carlo, z wykorzystaniem wielokrotnego powtarzania próbkowania populacji ogólnej. Wymienić tu należy przede wszystkim metody oparte na tzw. bootstrapie (Efron, Tibshirani 1993; Dybowski, Roberts 1999; Heskes 1997; Tibshirani 1996).

3. Podejścia bayesowskie, oparte na technikach wnioskowania bayesowskiego zastosowanych do zagadnień uczenia modelu neuronowego (neuronowo-rozmytego) oraz jego analizy i wnioskowania np. o rozkładzie wyjścia (Bishop 1995; MacKay 1991, 1994).

W naszej pracy pomijamy metody z trzeciej grupy – oparte na podejściu bayesowskim. Odnośnie do interesującego nas zakresu tematycznego teoria uczenia bayesowskiego daje w zasadzie alternatywne podstawy dla wyników otrzymywanych na gruncie zasady maksymalnej wiarygodności, na których bazują nasze rozważania w bieżącym rozdziale. Daje ona, co prawda, pewne punkty wyjścia do rozwoju metod omijających niektóre z założeń, jakie poczyniliśmy, ale w obecnej chwili mają one znaczenie raczej teoretyczne. Ewentualne ich zastosowanie wymaga bardzo kosztownego obliczeniowo wyznaczania całek w wielowymiarowych przestrzeniach metodą Monte Carlo.

W bieżącym punkcie przedstawimy więc szereg metod szacowania wariancji rozkładu prognozy spowodowanej niepewnością parametrów – wszystkie one pochodzą z dwóch pierwszych grup. Różnią się między sobą zarówno dokładnością otrzymywanych wyników, jak i efektywnością. W związku z tym przedstawimy również wyniki prac empirycznych wskazujących na przydatność poszczególnych rozwiązań w zagadnieniach krótkoterminowej prognozy zapotrzebowania na energię.

3.3.2. Metoda delta

Metoda delta polega na lokalnej linearyzacji sieci neuronowej (neuronowo-rozmytej) względem parametrów (wag) sieci, przy wykorzystaniu rozwinięcia Taylora pierwszego rzędu. Dzięki temu, jak zobaczymy, do oszacowania warunkowego rozkładu wyjściowego będziemy mogli wykorzystać wiele istotnych faktów, o których była mowa w punkcie 3.2.4, dotyczących modeli liniowych. W tym przypadku będą one miały oczywiście charakter przybliżony, a jakość przybliżenia zależy od tego, jak dobre jest założone przybliżenie modelu. Musimy jeszcze założyć, że spełnione są pewne warunki przyjęte w punkcie 3.2.4, związane z rozkładem błędu losowego $N(0, \sigma_\varepsilon)$ i niezależnością poszczególnych wartości odchyleń resztowych modelu otrzymanych dla obserwacji ze zbioru treningowego.

Przyjmijmy jak poprzednio, że mamy pewien zbiór treningowy D składający się ze wzorców wejściowych oraz odpowiadających im znanych (treningowych) wartości zmiennej wyjściowej: $D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$. Jeśli więc $f(\mathbf{x}, \mathbf{w}^*)$ jest siecią neuronową (neuronowo-rozmytą), nauczoną na zbiorze D metodą najmniejszych kwadratów, zaś $\mathbf{w}^*$ jest zbiorem parametrów (wag) sieci otrzymanym w procesie minimalizacji błędu kwadratowego

$$E = \frac{1}{2}\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}))^2 \tag{3.3.1}$$

to w dostatecznie małym otoczeniu punktu (wektora optymalnych wag) $\mathbf{w}^*$ możemy równanie modelu aproksymować rozwinięciem w szereg Taylora pierwszego rzędu:

$$f(\mathbf{x}, \mathbf{w}) = f(\mathbf{x}, \mathbf{w}^*) + \mathbf{g}(\mathbf{x}, \mathbf{w}^*)\Delta\mathbf{w} \tag{3.3.2}$$

gdzie $\mathbf{g}(\mathbf{x}, \mathbf{w}^*)$ jest wektorem gradientu wyjścia sieci względem wag $\mathbf{w}$, dla wartości optymalnej wag $\mathbf{w}^*$, zaś $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}^*$.

Jeżeli powyższa aproksymacja będzie dostatecznie dobra, to do modelu (3.2.3) możemy zastosować wiedzę przedstawioną w punkcie 3.2.4 odnośnie do liniowych modeli regresji. Przede wszystkim jeżeli model jest nieobciążony, to z założenia o rozkładzie błędu $N(0, \sigma_\varepsilon)$ wynika, że powodowany przez niego wspólny rozkład prawdopodobieństwa pomierzonych wartości treningowych y_i, dla poszczególnych $\mathbf{x}_k$, $k = 1, \ldots, N$, też będzie miał charakter rozkładu normalnego, a także – w konsekwencji – wspólny rozkład prawdopodobieństwa parametrów modelu minimalizujących błąd kwadratowy dla różnych realizacji próby (zbioru treningowego) jest również wielowymiarowym rozkładem normalnym $N(\mathbf{w}^*, \mathbf{C}_\mathbf{w})$:

$$p(\mathbf{w}) = k \exp(-\frac{1}{2}\Delta\mathbf{w}^{\mathrm{T}}\mathbf{C}_{\mathbf{w}}^{-1}\Delta\mathbf{w}) \tag{3.3.3}$$

gdzie wartością oczekiwaną rozkładu jest optymalny zestaw współczynników wagowych $\mathbf{w}^*$, otrzymany w procesie uczenia sieci, zaś $\mathbf{C}_\mathbf{w}$ jest macierzą kowariancji wag.

A oto dalsze konsekwencje przyjętej aproksymacji: ponieważ zależność między wagami a wyjściem modelu jest liniowa, to niepewność wyjścia modelu, wynikająca ze skończonego charakteru próby i powstającej w efekcie niepewności współczynników wagowych sieci, opisana jest również rozkładem normalnym $N(f(\mathbf{x}, \mathbf{w}^*), \sigma_\mathbf{w}(\mathbf{x}))$. W związku z tym warunkowy rozkład prognozy dla danego wejścia modelu, jako wynik złożenia dwóch niezależnych źródeł niepewności o rozkładzie normalnym, będzie miał również charakter rozkładu normalnego $N(f(\mathbf{x}, \mathbf{w}^*), \sigma_y(\mathbf{x}))$. Ewentualną niepewność wartości zmiennych wejściowych modelu będziemy na razie pomijać (problemem tym zajmiemy się w podrozdziale 3.5). Wariancja wyjściowa modelu $\sigma^2_y(\mathbf{x})$ określona jest zatem przez:

$$\sigma_y^2(\mathbf{x}) = \sigma_w^2(\mathbf{x}) + \sigma_\varepsilon^2 \tag{3.3.4}$$

Kluczową rolę w wyznaczeniu wariancji wyjściowej sieci neuronowej (neuronowo-rozmytej) spowodowanej niepewnością współczynników wagowych sieci $\sigma^2_w(\mathbf{x})$ będzie miało znalezienie macierzy kowariancji wag $\mathbf{C}_\mathbf{w}$. Dla zlinearyzowanej sieci (3.3.2) możemy wówczas zastosować wzór na wyznaczenie wariancji wyjściowej z parametrów modelu liniowego (3.2.32), wynikający z prawa propagacji błędów:

$$\sigma_w^2(\mathbf{x}) = \mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} \mathbf{C}_\mathbf{w} \mathbf{g}(\mathbf{x}, \mathbf{w}^*) \tag{3.3.5}$$

Do wyznaczenia macierzy $\mathbf{C}_\mathbf{w}$ podejdziemy nieco odmiennie niż w przypadku liniowym. Oczywiście moglibyśmy wyznaczyć macierz kowariancji w pełni za pomocą dostosowanego do nowej sytuacji wzoru (3.2.32) (a nawet, jak zobaczymy dalej, będzie to jedną z rozważanych możliwości), ale pamiętać należy, że wzór ten dla modeli liniowych spełniany jest dokładnie. Dla sieci neuronowych czy neuronowo-rozmytych, o nieliniowej charakterystyce, wyznaczona za pomocą (3.2.32) macierz kowariancji wag $\mathbf{C}_\mathbf{w}$ byłaby kolejnym przybliżeniem. Zobaczmy, czy nie uda się nam znaleźć jej lepszego oszacowania. Spróbujemy je uzyskać, wykorzystując znane w teorii optymalizacji nieliniowej podejście oparte na aproksymacji kwadratowej funkcji błędu.

Przypomnijmy sobie dyskusję, jaką prowadziliśmy na koniec punktu 3.2.4, dotyczącą zależności między zmianami wartości błędu kwadratowego w funkcji

parametrów modelu a elipsoidami ufności i kowariancją określającą niepewność parametrów. Dla dowolnego zestawu wag **w** i wag optymalnych $\mathbf{w}^*$ minimalizujących błąd kwadratowy sieci na zbiorze treningowym na podstawie zależności (3.2.44) możemy dla naszego przybliżonego modelu (3.3.2) wyznaczyć różnicę błędu kwadratowego:

$$E(\mathbf{w}) - E(\mathbf{w}^*) = \frac{1}{2\sigma_\varepsilon^2} \Delta\mathbf{w}^{\mathrm{T}} \mathbf{C}_{\mathbf{w}}^{-1} \Delta\mathbf{w} \tag{3.3.6}$$

gdzie, jak wcześniej, $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}^*$.

Z drugiej jednak strony, stosując w stosunku do funkcji błędu $E(\mathbf{w})$ aproksymację kwadratową, czyli przybliżając ją funkcją kwadratową poprzez rozwinięcie w szereg Taylora drugiego rzędu, otrzymujemy:

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \Delta\mathbf{w}^{\mathrm{T}} \nabla E(\mathbf{w}^*) + \frac{1}{2} \Delta\mathbf{w}^{\mathrm{T}} \nabla\nabla E(\mathbf{w}^*) \Delta\mathbf{w} \tag{3.3.7}$$

gdzie $\nabla E(\mathbf{w}^*)$ jest gradientem błędu E względem wag sieci w punkcie $\mathbf{w}^*$, zaś $\nabla\nabla E(\mathbf{w}^*)$ jego hesjanem. Zauważmy, że w punkcie minimum $\mathbf{w}^*$ gradient błędu, jako wektor pochodnych cząstkowych względem poszczególnych wag, jest równy 0. Oznaczmy ponadto hesjan $\nabla\nabla E(\mathbf{w}^*)$ (macierz drugich pochodnych względem wag) przez **H**. Wówczas (3.3.7) możemy zapisać w następujący sposób:

$$E(\mathbf{w}) - E(\mathbf{w}^*) = \frac{1}{2} \Delta\mathbf{w}^{\mathrm{T}} \mathbf{H} \Delta\mathbf{w} \tag{3.3.8}$$

Jeżeli teraz porównamy zależności (3.3.6) i (3.3.8):

$$\frac{1}{2\sigma_\varepsilon^2} \Delta\mathbf{w}^{\mathrm{T}} \mathbf{C}_w^{-1} \Delta\mathbf{w} = E(\mathbf{w}) - E(\mathbf{w}^*) = \frac{1}{2} \Delta\mathbf{w}^{\mathrm{T}} \mathbf{H} \Delta\mathbf{w} \tag{3.3.9}$$

to widzimy natychmiast, że macierz kowariancji $\mathbf{C}_\mathbf{w}$ wag sieci neuronowej (neuronowo-rozmytej) możemy wyznaczyć na podstawie macierzy **H**, hesjanu błędu względem parametrów modelu (wag sieci neuronowej lub neuronowo-rozmytej), mnożąc jej odwrotność przez wariancję czynnika losowego:

$$\mathbf{C}_\mathbf{w} = \sigma_\varepsilon^2 \mathbf{H}^{-1} \tag{3.3.10}$$

Zauważmy przy tym, że macierz odwrotności hesjanu $\mathbf{H}^{-1}$ spełnia także inne warunki, których wymagamy od macierzy kowariancji. Jeśli sieć ma p współczynników wagowych, to $\mathbf{H}$ jest macierzą kwadratową $p \times p$, ponadto jest to macierz symetryczna. Oczywiście wiadomo również, że w punkcie minimum błędu jego hesjan (jako macierz drugich pochodnych) jest dodatnio określony. Wobec tego symetryczna i dodatnio określona będzie również macierz odwrotna $\mathbf{H}^{-1}$.

Zastępując więc we wzorze (3.3.5) macierz kowariancji współczynników wagowych sieci macierzą $\mathbf{C}_\mathbf{w}$ wyznaczoną przez (3.3.10), otrzymujemy oszacowanie wariancji wyjściowej modelu wywoływanej przez niepewność tych parametrów:

$$\sigma_{\mathbf{w}}^2(\mathbf{x}) = \sigma_\varepsilon^2 \mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{g}(\mathbf{x}, \mathbf{w}^*) \tag{3.3.11}$$

oraz łącznej wariancji wyjściowej modelu, czyli warunkowego rozkładu prawdopodobieństwa prognozy dla danego wzorca wejściowego:

$$\sigma_y^2(\mathbf{x}) = \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{g}(\mathbf{x}, \mathbf{w}^*) \tag{3.3.12}$$

Wyznaczenie hesjanu funkcji względem jej argumentów lub parametrów stanowi w wielu przypadkach podstawę algorytmów optymalizacji wielowymiarowej. Dlatego też w literaturze poświęconej zagadnieniom metod numerycznych znaleźć można wiele różnorodnych podejść do rozwiązania tego problemu. Obszerny przegląd tych zagadnień zaprezentowany został np. w publikacji: Press, Teukolsky, Vetterling, Flannery 1992. Ostateczna postać macierzy hesjanu zależna jest oczywiście od konkretnej budowy minimalizowanej funkcji, a więc w naszym przypadku od struktury modelu.

Przeanalizujemy przy tym dwa najważniejsze podejścia do wyznaczania hesjanu błędu kwadratowego dla nieliniowych metod szacowania funkcji regresji, takich jak sieci neuronowe czy neuronowo-rozmyte. Pierwsze z nich ma charakter przybliżony i polega na aproksymacji hesjanu z wykorzystaniem pierwszych pochodnych wyjścia sieci względem współczynników wagowych, wyznaczonych na całym zbiorze treningowym, natomiast drugie związane jest z dokładnym wyznaczeniem hesjanu sieci.

Niezależnie od struktury wewnętrznej modelu, wyznaczmy pochodną kwadratu odchylenia modelu (reszty) e^2_k, dla dowolnej obserwacji ze zbioru treningowego (próby), względem dowolnego parametru modelu (wagi sieci) w_i, $k = 1, \ldots, N$, $i = 1, \ldots, p$. Przez N rozumiemy liczbę obserwacji w zbiorze treningowym, przez p liczbę parametrów (wag) modelu. Korzystając ze wzoru na pochodną funkcji złożonej, natychmiast otrzymujemy następującą zależność:

$$\frac{\partial e_k^2}{\partial w_i} = \frac{\partial (y_k - f(\mathbf{x}_k, \mathbf{w}))^2}{\partial w_i}\bigg|_{\mathbf{w}=\mathbf{w}^*} = -2(y_k - f(\mathbf{x}_k, \mathbf{w}))\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_i}\bigg|_{\mathbf{w}=\mathbf{w}^*} \tag{3.3.13}$$

Pochodną $e^2{}_k$ względem w_i obliczamy naturalnie w punkcie $\mathbf{w}^*$, tj. w punkcie optymalnych wartości wag sieci wyznaczonych w procesie treningu. Stosując dalej wzór na pochodną iloczynu funkcji, możemy wyznaczyć drugą pochodną odchylenia $e^2{}_k$ względem wag sieci:

$$\frac{\partial^2 e_k^2}{\partial w_i \partial w_j} = 2\left(\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_i}\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_j} - (y_k - f(\mathbf{x}_k, \mathbf{w}))\frac{\partial^2 f(\mathbf{x}_k, \mathbf{w})}{\partial w_i \partial w_j}\right)\bigg|_{\mathbf{w}=\mathbf{w}^*} \tag{3.3.14}$$

Druga pochodna całości błędu kwadratowego E (zależność (3.3.1)) względem parametrów modelu będzie sumą członów (3.3.14) obliczonych dla wszystkich obserwacji ze zbioru treningowego. Elementy macierzy hesjanu h_{ij}, i, $j = 1, \ldots, p$ możemy zatem wyznaczyć następująco:

$$\begin{aligned} h_{ij} &= \frac{\partial^2 E}{\partial w_i \partial w_j}\bigg|_{\mathbf{w}=\mathbf{w}^*} = \frac{1}{2}\sum_{k=1}^{N}\frac{\partial^2 e_k^2}{\partial w_i \partial w_j}\bigg|_{\mathbf{w}=\mathbf{w}^*} = \\ &= \sum_{k=1}^{N}\left(\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_i}\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_j} - (y_k - f(\mathbf{x}_k, \mathbf{w}))\frac{\partial^2 f(\mathbf{x}_k, \mathbf{w})}{\partial w_i \partial w_j}\right)\bigg|_{\mathbf{w}=\mathbf{w}^*} \end{aligned} \tag{3.3.15}$$

Jeżeli przyjrzymy się dokładniej zależności (3.3.15), to widzimy, że każdy element macierzy hesjanu błędu kwadratowego **H** składa się z dwóch członów: pierwszego, opartego na pierwszych pochodnych (gradientach) wyjścia modelu względem parametrów, i drugiego – opartego na drugich pochodnych:

$$h_{ij} = \sum_{k=1}^{N}\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_i}\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_j}\bigg|_{\mathbf{w}=\mathbf{w}^*} - \sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}))\frac{\partial^2 f(\mathbf{x}_k, \mathbf{w})}{\partial w_i \partial w_j}\bigg|_{\mathbf{w}=\mathbf{w}^*} \tag{3.3.16}$$

Zauważmy dalej, że drugi składnik w (3.3.16) powinien być wielkością małą, bliską zera, tak więc, jeżeli go zaniedbamy, i do obliczenia elementów hesjanu **H** wykorzystamy tylko pierwszy z nich, to błąd przybliżenia powinien być nieduży. Jeżeli bowiem linearyzacja sieci neuronowej (neuronowo-rozmytej) (3.3.2) względem parametrów jest dostatecznie dobra, to drugie pochodne sieci względem tych parametrów powinny zanikać albo przynajmniej być bardzo małe w porównaniu z członami gradientów. Rzecz jasna, druga pochodna funkcji liniowej jest równa 0. Ponadto w (3.3.16) pochodne te

mnożone są przez reszty sieci i sumowane po całym zbiorze treningowym. Przypomnijmy, że zgodnie z przyjętymi założeniami, reszty modelu są niezależne i mają symetryczny rozkład normalny o wartości oczekiwanej 0. Powinny więc być one równomiernie rozłożone względem 0. W związku z tym i tak już raczej niewielkie człony zawierające drugą pochodną będą mnożone naprzemiennie przez liczby dodatnie oraz ujemne, co dodatkowo spowoduje ich wzajemną (oczywiście częściową) redukcję podczas sumowania.

Podsumowując, ponieważ drugi składnik w (3.2.16) powinien mieć dużo mniejszą wartość niż pierwszy, możemy spróbować go zaniedbać i aproksymować elementy macierzy hesjanu błędu **H** za pomocą pierwszego z nich:

$$h_{ij} = \sum_{k=1}^{N} \frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_i} \frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_j}\bigg|_{\mathbf{w}=\mathbf{w}^*} \tag{3.3.17}$$

Zdefiniujmy macierz $\mathbf{\Phi}$, o wymiarach $N \times p$, której wiersze odpowiadają poszczególnym obserwacjom zbioru treningowego i składają się z gradientów wyjścia sieci względem poszczególnych parametrów wagowych w_i, wyznaczonych dla wzorca wejściowego danej obserwacji treningowej $\mathbf{x}_k$:

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{g}(\mathbf{x}_1, \mathbf{w}^*)^{\mathrm{T}} \\ \mathbf{g}(\mathbf{x}_2, \mathbf{w}^*)^{\mathrm{T}} \\ \cdots \\ \mathbf{g}(\mathbf{x}_N, \mathbf{w}^*)^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f(\mathbf{x}_1, \mathbf{w})}{\partial w_1} & \frac{\partial f(\mathbf{x}_1, \mathbf{w})}{\partial w_2} & \cdots & \frac{\partial f(\mathbf{x}_1, \mathbf{w})}{\partial w_p} \\ \frac{\partial f(\mathbf{x}_2, \mathbf{w})}{\partial w_1} & \frac{\partial f(\mathbf{x}_2, \mathbf{w})}{\partial w_2} & \cdots & \frac{\partial f(\mathbf{x}_2, \mathbf{w})}{\partial w_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(\mathbf{x}_N, \mathbf{w})}{\partial w_1} & \frac{\partial f(\mathbf{x}_N, \mathbf{w})}{\partial w_2} & \cdots & \frac{\partial f(\mathbf{x}_N, \mathbf{w})}{\partial w_p} \end{bmatrix}_{\mathbf{w}=\mathbf{w}^*} \tag{3.3.18}$$

Wówczas zależność wyznaczającą macierz hesjanu błędu względem wag (3.3.17) możemy zapisać następująco:

$$\mathbf{H} = \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi} \tag{3.3.19}$$

Ponieważ we wzorze (3.3.19) mamy do czynienia z iloczynem skalarnym kolumn macierzy $\mathbf{\Phi}$ (3.3.18) przez siebie, to powyższe oszacowanie macierzy hesjanu błędu modelu określa się zazwyczaj jako tzw. aproksymację iloczynem skalarnym (*outer product approximation*), czasami również nazywaną aproksymacją Levenberga–Marquardta (Bishop 1995). Zastępując w (3.3.12) macierz **H** powyższym przybliżeniem, otrzymamy oszacowanie wariancji wyjściowej modelu:

$$\sigma_y^2(\mathbf{x}) = \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} (\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi})^{-1} \mathbf{g}(\mathbf{x}, \mathbf{w}^*) \tag{3.3.20}$$

Zauważmy, że wykorzystanie aproksymacji iloczynem skalarnym oznacza w zasadzie przyjęcie założenia zupełnej linearyzacji modelu. Jak już wspomnieliśmy, człony związane z drugimi pochodnymi modelu względem parametrów w (3.3.16) zanikają w przypadku modelu liniowego. Ponadto jeśli spojrzymy na linearyzację modelu we wzorze (3.3.2), $f(\mathbf{x}, \mathbf{w}) - f(\mathbf{x}, \mathbf{w}^*) = \mathbf{g}(\mathbf{x}, \mathbf{w}^*)\Delta\mathbf{w}$ i potraktujemy go jako model liniowy, którego wejściami są wartości gradientów sieci $\mathbf{g}(\mathbf{x}, \mathbf{w}^*)$, zaś parametrami zmiany wag $\Delta\mathbf{w}$, to zauważymy, że macierz $\mathbf{\Phi}$ jest dla tego modelu dokładnie odpowiednikiem macierzy $\mathbf{X}$ z układu równań (3.2.17) dla liniowego modelu regresji, zaś wzór (3.3.20) dokładnie odpowiednikiem wzoru (3.2.33) na oszacowanie wariancji wyjściowej modelu liniowego.

Do określenia elementów macierzy $\mathbf{\Phi}$ (3.3.18) niezbędne jest wyznaczenie dla wszystkich wzorców w zbiorze treningowym pochodnych wyjścia sieci neuronowej (neuronowo-rozmytej) względem współczynników wagowych. Podobnie, aby zastosować wzór (3.3.20), musimy wyznaczyć również $\mathbf{g}(\mathbf{x}, \mathbf{w}^*)$, gradient wyjścia sieci względem wag, dla nowego wzorca wejściowego $\mathbf{x}$. Odpowiednie algorytmy pozwalające na ich obliczenie dla najważniejszych wykorzystywanych w pracy modeli sieci neuronowych i neuronowo-rozmytych przedstawione zostały w załącznikach. Dla warstwowych sieci perceptronowych MLP zagadnienie wyznaczania pochodnych wyjścia sieci względem wag zaprezentowano w załączniku 1 w punkcie Z1.3. Analogicznie dla sieci neuronowo-rozmytych FBF odpowiednie zależności dla elementów gradientu sieci w przestrzeni wag przedstawione zostały załączniku 2 w punkcie Z2.3, zaś dla sieci neuronowo-rozmytych implementujących wnioskowanie rozmyte typu Takagi–Sugeno, z liniowymi następnikami reguł, w załączniku 3 w punkcie Z3.3.

Przybliżenie hesjanu błędu przy użyciu aproksymacji iloczynem skalarnym daje zazwyczaj, z praktycznego punktu widzenia, wystarczająco dokładne wyniki oszacowania wariancji rozkładu prognozy (Chryssolouris, Lee, Ramsey 1996). MacKay pokazuje jednak, że w pewnych przypadkach weryfikacje eksperymentalne tego podejścia mogą być niesatysfakcjonujące (MacKay 1991, 1994). W sytuacji, w której drugi składnik (3.3.16) jest na tyle duży, że nie może zostać zaniedbany, alternatywnym rozwiązaniem może być dokładne wyznaczenie macierzy $\mathbf{H}$, z pełnym oszacowaniem również członów drugich pochodnych. Algorytmy obliczeń dokładnych wartości hesjanu błędu modelu względem wag są nieco bardziej złożone i wymagają większych nakładów obliczeniowych niż w przypadku jego aproksymacji iloczynem skalarnym, mogą one jednak dawać wyraźnie dokładniejsze oszacowania wariancji wyjściowej modelu. Odpowiednie równania dla badanych w pracy modeli również zaprezentowane zostały w załącznikach. Algorytm wyznaczania drugich pochodnych błędu dla warstwowej sieci perceptronowej przedstawiono w załączniku 1 w punkcie Z1.2, dla

sieci neuronowo-rozmytej FBF – w załączniku 2 w punkcie Z2.2, zaś dla sieci neuronowo-rozmytych implementujących wnioskowanie rozmyte typu Takagi–Sugeno, z liniowymi następnikami reguł – w załączniku 3 w punkcie Z3.2.

Zanim przejdziemy do omówienia wyników zastosowań prezentowanych tutaj metod szacowania wariancji wyjściowej modelu w zagadnieniach prognozy zapotrzebowania na energię elektryczną, musimy zwrócić uwagę jeszcze na jedną kwestię. W prezentowanych dotąd rozważaniach przyjmuje się, że rozkład reszt modelu, a co za tym idzie, błąd losowy modelu ma charakter rozkładu normalnego $N(0, \sigma_\varepsilon)$, o wartości oczekiwanej 0 i stałym odchyleniu standardowym σ_ε. W konsekwencji w zależnościach na wariancję wyjściową prognozy, zarówno w przypadku zastosowania aproksymacji iloczynem skalarnym (3.3.20), jak i w ogólniejszym przypadku pełnego hesjanu błędu (3.3.12), występuje stała wariancja błędu losowego σ_ε^2. W wielu jednak przypadkach, zwłaszcza dla nieliniowych modeli prognostycznych, mamy do czynienia z tak zwaną heteroskedastycznością, to znaczy sytuacją, w której reszty (błędy losowe) modelu podlegają rozkładom normalnym, ale o zmiennych odchyleniach standardowych $\sigma_\varepsilon(\mathbf{x})$, zależnych od wzorca wejściowego prognozy $\mathbf{x}$.

Problemem tym dokładniej zajmiemy się w następnym podrozdziale 3.4, ale z przeprowadzonych przez nas badań wynika, że właśnie z taką sytuacją mieliśmy do czynienia podczas naszych doświadczeń w zakresie prognozowania krótkoterminowego zapotrzebowania na energię elektryczną. Wymaga to dostarczenia odpowiedniego oszacowania odchylenia standardowego (lub wariancji) czynnika losowego $\sigma_\varepsilon(\mathbf{x})$, dla dowolnej wartości wejściowej $\mathbf{x}$, co będzie jednym z głównych punktów dyskusji w podrozdziale 3.4. Zauważmy jednak, że ma to również pewien wpływ na wyznaczenie macierzy kowariancji wag i wynikającej z niej wariancji wyjściowej modelu.

W przypadku występowania błędu losowego o wariancji zależnej od wartości zmiennych wejściowych, podobnie jak już to omawialiśmy w punkcie 3.2.4 dla przypadku modeli liniowych, powinniśmy do uczenia modelu zastosować metodę ważonych najmniejszych kwadratów minimalizującą ważony błąd kwadratowy, w którym każde odchylenie prognozy normalizowane jest przez błąd poszczególnych próbek treningowych:

$$E = \frac{1}{2}\sum_{k=1}^{N}\frac{(y_k - f(\mathbf{x}_k, \mathbf{w}))^2}{\sigma_\varepsilon^2(\mathbf{x}_k)} \tag{3.3.21}$$

Elementy macierzy hesjanu $\mathbf{H}$, ważonego błędu kwadratowego, możemy wyznaczyć na podstawie zależności:

$$h_{ij} = \left.\frac{\partial^2 E}{\partial w_i \partial w_j}\right|_{\mathbf{w}=\mathbf{w}^*} = \sum_{k=1}^{N}\frac{1}{\sigma_\varepsilon^2(\mathbf{x}_k)}\frac{\partial^2}{\partial w_i \partial w_j}\left(\frac{1}{2}e_k^2\right)\Bigg|_{\mathbf{w}=\mathbf{w}^*} \tag{3.3.22}$$

Jak więc widzimy, w przypadku dopasowania modelu metodą ważonych najmniejszych kwadratów procedury obliczeniowe elementów hesjanu błędu będą niemal identyczne jak przy użyciu metody zwyczajnych normalnych kwadratów. Zarówno w przypadku aproksymacji iloczynem skalarnym (3.3.19), jak i algorytmów pełnego oszacowania hesjanu dla wybranych typów modeli neuronowych i neuronowo-rozmytych prezentowanych w załącznikach w punktach Z1.2, Z2.2 i Z.2, wystarczy każdy człon hesjanu wyznaczony dla pojedynczego wzorca treningowego przed zsumowaniem podzielić przez $\sigma_\varepsilon^2(\mathbf{x}_k)$.

Oczywiście w przypadku metody ważonych najmniejszych kwadratów macierz kowariancji wag modelu jest równa:

$$\mathbf{C}_\mathbf{w} = \mathbf{H}^{-1} \tag{3.3.23}$$

Zauważmy, że w przeciwieństwie do wzoru (3.3.10) dla zwyczajnych najmniejszych kwadratów nie mnożymy odwrotności hesjanu błędu **H** przez wariancję błędu losowego. Jej wartość została już uwzględniona wcześniej podczas obliczeń samego hesjanu.

Dopasowanie modelu do danych treningowych metodą ważonych najmniejszych kwadratów wymaga określenia wariancji błędów losowych poszczególnych wzorców z góry, przed rozpoczęciem procesu uczenia sieci. W praktyce jednak na etapie przygotowywania predyktora informacje te zazwyczaj są niedostępne. W związku z tym w rzeczywistych zastosowaniach prognostycznych modele na ogół dopasowuje się metodą zwyczajnych najmniejszych kwadratów, wykorzystując błąd kwadratowy SSE, czyli sumę kwadratów (3.3.1).

Również w przypadku rozważanych przez nas prognoz krótkoterminowych zapotrzebowania na energię elektryczną otrzymujemy omawiany w podrozdziale 3.4 estymator odchylenia standardowego błędu, zależny od wartości zmiennych wejściowych, ale dla wcześniej dopasowanego modelu prognostycznego, po zakończeniu procesu jego treningu. Do samego uczenia omawianych w pracy modeli neuronowo-rozmytych stosujemy metodę zwyczajnych najmniejszych kwadratów. W związku z tym oszacowania wariancji rozkładu prognozy wykonujemy na bazie teorii związanej z tą metodą, wykorzystując zależności omawiane w bieżącym punkcie. Zastępujemy w nich jedynie stałe odchylenie standardowe błędu losowego σ_ε wartością zależną od danego wzorca wejściowego prognozy $\sigma_\varepsilon(\mathbf{x})$. Oczywiście pamiętać należy, że takie postępowanie stanowi kolejne przybliżenie, które musi być zweryfikowane empirycznie.

Podsumujmy więc nasze rozważania w bieżącym punkcie. Prezentowana w nim metoda delta szacowania wartości wariancji wyjściowej prognozy, dla danego wzorca wejściowego **x**, przyjmuje następującą postać (zastępując we wzorze (3.3.12) stałe odchylenie standardowe σ_ε przez zależne od wejścia $\sigma_\varepsilon(\mathbf{x})$):

$$\sigma_y^2(\mathbf{x}) = \sigma_\varepsilon^2(\mathbf{x}) + \sigma_\varepsilon^2(\mathbf{x})\mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} \mathbf{H}^{-1}\mathbf{g}(\mathbf{x}, \mathbf{w}^*) \tag{3.3.24}$$

gdzie $\sigma_\varepsilon^2(\mathbf{x})$ jest wariancją błędu losowego, dla danego wzorca wejściowego prognozy $\mathbf{x}$, której modelowaniem zajmiemy się w punkcie 3.4. Wartość $\mathbf{g}(\mathbf{x}, \mathbf{w}^*)$ jest gradientem wyjścia sieci względem wag, dla tej prognozy. Metody wyznaczania gradientów dla wykorzystywanych w pracy modeli neuronowo-rozmytych, jak już wspomnieliśmy, znaleźć można w załącznikach Z1 – Z3. Tam przedstawione zostały również algorytmy dokładnego wyznaczania hesjanu błędu kwadratowego tych modeli neuronowo-rozmytych, które mogą zostać zastosowane do obliczenia macierzy $\mathbf{H}$ w formule (3.3.24).

Zamiast dokładnych obliczeń pełnego hesjanu błędu $\mathbf{H}$, możemy próbować zastosować jego aproksymację metodą iloczynu skalarnego (Levenberga–Marquarda), co daje w wyniku formułę na oszacowanie wariancji prognozy, dla danego wzorca wejściowego, analogiczną do (3.3.20). Jeżeli we wzorze (3.3.20) zastąpimy stałe odchylenie standardowe σ_ε przez $\sigma_\varepsilon(\mathbf{x})$, otrzymamy następującą zależność:

$$\sigma_y^2(\mathbf{x}) = \sigma_\varepsilon^2(\mathbf{x}) + \sigma_\varepsilon^2(\mathbf{x})\mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{g}(\mathbf{x}, \mathbf{w}^*) \tag{3.3.25}$$

Przypomnijmy, że $\mathbf{\Phi}$ jest macierzą daną przez (3.3.18). Jej wiersze składają się z gradientów wyjścia modelu względem parametrów wagowych $\mathbf{g}(\mathbf{x}_k, \mathbf{w}^*)$ wyznaczonych dla wzorca wejściowego danej obserwacji treningowej $\mathbf{x}_k$.

Aby ocenić możliwość wykorzystania metody delta do oszacowania wariancji (odchylenia standardowego) prognozy obciążenia sieci elektroenergetycznej, wykonano szereg badań symulacyjnych i zweryfikowano działanie metody na drodze empirycznej. Przedstawione wcześniej zależności wykorzystano do wyznaczenia przedziałów prognozy w przypadku szeregu zadań prognostycznych omawianych w rozdziale 2. Przypomnijmy, że przy przyjętych założeniach rozkład prognozy będzie miał charakter rozkładu normalnego $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$:

$$p(y/\mathbf{x}) = \frac{1}{\sigma_y(\mathbf{x})\sqrt{2\pi}} \exp\left(-\frac{(y - f(\mathbf{x}, \mathbf{w}))^2}{2\sigma_y^2(\mathbf{x})}\right) \tag{3.3.26}$$

o wartości oczekiwanej określonej przez wyjście modelu neuronowego lub neuronowo-rozmytego (prognozę) i odchyleniu standardowym $\sigma_y(\mathbf{x})$.

W naszych badaniach podczas tworzenia modeli wykorzystywaliśmy duże próby treningowe, złożone z kilkuset elementów. W takich przypadkach przedział wartości prognozy dla konkretnego wzorca wejściowego $\mathbf{x}$, przy przyjętym poziomie prawdopodobieństwa α, możemy wyznaczyć, korzystając

bezpośrednio z kwantyli rozkładu normalnego. Dolny i górny kraniec takiego przedziału będą równe:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= Q_{N\left(f(\mathbf{x},\mathbf{w}),\sigma_y(\mathbf{x})\right)}((1-\alpha)/2) \\ g_{y/\mathbf{x}}(\alpha) &= Q_{N\left(f(\mathbf{x},\mathbf{w}),\sigma_y(\mathbf{x})\right)}((1+\alpha)/2) \end{aligned} \tag{3.3.27}$$

gdzie przez $Q(\alpha)$ oznaczyliśmy kwantyl warunkowego rozkładu prawdopodobieństwa prognozowanej wielkości (3.3.26) dla prawdopodobieństwa α. Alternatywnie, dokładnie ten sam przedział prognozy możemy wyznaczyć, korzystając z kwantyli rozkładu normalnego standardowego $N(0, 1)$, $Q_{N(0,1)}(\alpha)$:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) - Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_y(\mathbf{x}) \\ g_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) + Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_y(\mathbf{x}) \end{aligned} \tag{3.3.28}$$

Zauważmy, że w (3.3.28) przy wyznaczaniu dolnego krańca przedziału prognozy korzystamy z symetrii rozkładu normalnego standardowego, wokół wartości oczekiwanej 0. Z symetrii tej wynika zależność $Q_{N(0,1)}((1-\alpha)/2) = -Q_{N(0,1)}((1+\alpha)/2)$. Przypomnijmy jeszcze, że jeżeli zbiór treningowy, na podstawie którego wyznaczamy przecież również oszacowanie odchylenia standardowego (wariancji) prognozy $\sigma_y(\mathbf{x})$, zawiera mniej wzorców, należy zastąpić w (3.3.28) kwantyl rozkładu normalnego kwantylem rozkładu t-Studenta:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) - t_{N-p}((1+\alpha)/2)\cdot\sigma_y(\mathbf{x}) \\ g_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) + t_{N-p}((1+\alpha)/2)\cdot\sigma_y(\mathbf{x}) \end{aligned} \tag{3.3.29}$$

gdzie $t_{N-p}((1+\alpha)/2)$ jest jednostronnym kwantylem rozkładu t-Studenta, dla prawdopodobieństwa $(1+\alpha)/2$ i $N–p$ stopni swobody. Przez N rozumiemy (jak zwykle w naszej pracy) liczbę wzorców treningowych, natomiast p oznacza tutaj liczbę parametrów (wag) modelu neuronowego lub neuronowo-rozmytego. Dokładniej rzecz biorąc, wartość p określać powinna liczbę tzw. parametrów efektywnych sieci (problem ten przedstawiać będziemy bliżej w punkcie 3.4).

Analizując przedziały prognozy otrzymane dla badanych modeli, we wszystkich przypadkach porównujemy wyniki przy użyciu do oszacowania odchylenia standardowego metodą wyznaczenia pełnego hesjanu błędu kwadratowego:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) - Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_\varepsilon(\mathbf{x})\sqrt{1+\mathbf{g}(\mathbf{x},\mathbf{w}^*)^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{g}(\mathbf{x},\mathbf{w}^*)} \\ g_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) + Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_\varepsilon(\mathbf{x})\sqrt{1+\mathbf{g}(\mathbf{x},\mathbf{w}^*)^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{g}(\mathbf{x},\mathbf{w}^*)} \end{aligned} \tag{3.3.30}$$

oraz metodą aproksymacji hesjanu za pomocą iloczynu skalarnego:

$$d_{y/\mathbf{x}}(\alpha) = f(\mathbf{x},\mathbf{w}) - Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_{\varepsilon}(\mathbf{x})\sqrt{1+\mathbf{g}(\mathbf{x},\mathbf{w}^*)^{\mathrm{T}}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{g}(\mathbf{x},\mathbf{w}^*)}$$
$$g_{y/\mathbf{x}}(\alpha) = f(\mathbf{x},\mathbf{w}) + Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_{\varepsilon}(\mathbf{x})\sqrt{1+\mathbf{g}(\mathbf{x},\mathbf{w}^*)^{\mathrm{T}}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{g}(\mathbf{x},\mathbf{w}^*)} \quad (3.3.31)$$

Jako pierwszy problem praktyczny, na którym zweryfikujemy doświadczalnie zastosowanie metody delta do wyznaczenia wariancji (odchylenia standardowego) warunkowego rozkładu prawdopodobieństwa prognozowanego zapotrzebowania na energię dla określonego wzorca wejściowego, rozważmy prezentowane w punktach 2.2.4 i 2.2.5 zagadnienie prognozy godzinnego zapotrzebowania na energię z dwudniowym wyprzedzeniem czasowym. Przypomnijmy, że model ten określony był równaniem:

$$ZG_d(t) = f(ZG(t)_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZG(t)_{d-2}, ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \quad (3.3.32)$$

gdzie oznaczenia są takie same jak w punkcie 2.2.4, tzn.:

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d,
$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,
dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Tabela 3.3.1. Częstości empiryczne przedziałów prognozy godzinnego zapotrzebowania na energię, otrzymanych za pomocą metody delta dla sieci MLP (w %)

Prawdopodobieństwo	Dokładny hesjan błędu				Przybliżony hesjan błędu			
	godz. 7	godz. 12	godz. 17	godz. 20	godz. 7	godz. 12	godz. 17	godz. 20
80	78,21	82,32	78,45	80,87	86,56	87,29	87,41	86,32
85	84,02	85,84	82,93	84,02	90,07	90,92	91,04	90,56
90	88,38	89,23	87,77	88,86	93,70	94,55	93,83	93,34
95	93,46	93,10	92,37	92,74	97,34	96,97	96,25	95,52

Źródło: opracowanie własne.

Do prognozy wykorzystano oddzielne sztuczne sieci neuronowe MLP (warstwowe sieci perceptronowe) o strukturze {13, 10, 1} (13 neuronów w warstwie wejściowej, 10 w warstwie ukrytej, 1 neuron wyjściowy). Dla powyższych modeli przetestowano przedziały prognozy wyznaczane za pomocą metody delta, dla pełnego hesjanu błędu (przy użyciu zależności (3.3.30)) oraz przybliżonego hesjanu (zależność (3.3.31)) (Bartkiewicz 1999b, c; Bartkiewicz, Gontar, Zieliński, Bardzki 2000b; Bartkiewicz 2000a, Bartkiewicz 2001a). W procesie testowania wykorzystano długi, niemal trzyletni zbiór danych, złożony z ponad ośmiuset wzorców testowych. Dane te obejmowały wszystkie rodzaje dni tygodnia, włączając w to również soboty i niedziele, odrzucono jednakże święta

narodowe i religijne (oraz dni występujące bezpośrednio po nich). Tego rodzaju dni nietypowe prognozowane były przy użyciu specjalnego podejścia przez odrębne modele (dokładniejsze wyjaśnienia znaleźć można w punkcie 2.2.5).

Wyniki testowania przedziałów prognozy w przypadku typowych poziomów prawdopodobieństwa, przeprowadzone dla modeli neuronowych sporządzonych dla czterech wybranych godzin, przedstawione zostały tabeli 3.3.1. Zawiera ona informacje o odsetku obserwacji rzeczywistego zapotrzebowania na energię w obrębie przedziałów wyznaczonych dla tych prognoz. Ponieważ prognozy dla poszczególnych godzin sporządzane są za pomocą oddzielnych modeli, to wyniki w tabeli 3.3.1 traktować możemy jako badania empiryczne przedziałów prognozy dla czterech niezależnych przypadków sieci MLP.

Jak widzimy, w przypadku oszacowania wariancji rozkładu prognozy zapotrzebowania na energię elektryczną za pomocą przybliżonej macierzy hesjanu błędu sieci neuronowej, obliczonej metodą aproksymacji iloczynem skalarnym (zależność (3.3.31)), dokładność wyznaczonych przedziałów prognozy jest wyraźnie gorsza. Jeżeli spojrzymy na drugą część tabeli 3.3.1, łatwo zauważymy, że uzyskane wartości częstości, czyli empiryczne prawdopodobieństwa znalezionych przedziałów, odbiegają jednak nieco od teoretycznego poziomu prawdopodobieństwa, dla którego przedziały te wyznaczono. Dokładność możemy ocenić jako względnie dobrą jedynie w samych ogonach rozkładu prawdopodobieństwa zapotrzebowania na energię. Dużo gorsze wyniki obserwujemy dla poziomów prawdopodobieństwa 80% i 85%.

Dużo lepszą dokładność uzyskano przy wykorzystaniu algorytmu wyznaczającego dokładny hesjan błędu sieci neuronowej MLP (zależność (3.3.30)). Empiryczne oszacowania prawdopodobieństwa wyznaczonych przedziałów prognozy są bardzo bliskie poziomom, dla których je wyznaczano. W najgorszym przypadku, to jest dla sieci neuronowej prognozującej zapotrzebowanie na energię o godzinie 17, dla 95% przedziału prognozy otrzymano 92,37% obserwacji wewnątrz oszacowanych przedziałów (763 wzorce testowe na łączną liczbę 826). Oznacza to jednak, że ryzyko konieczności znacznego rozszerzenia prawdopodobieństwa jest niewielkie. Na przykład prawdopodobieństwo, że prawdziwy poziom tego przedziału wynosi nie 95%, lecz tylko 90%, jest bardzo małe, ponieważ wynosi zaledwie około 1,1% (wynika to z wartości dystrybuanty rozkładu dwumianowego dla 63 sukcesów przy 826 próbach i prawdopodobieństwie sukcesu 0,9 – zagadnienia te omawiamy dokładniej w punkcie 3.2.2).

Drugim zadaniem prognostycznym z zakresu prognoz krótkoterminowych obciążeń sieci elektroenergetycznej, dla którego przeprowadziliśmy badania przedziałów prognozy, otrzymanych przy wykorzystaniu metody delta, był problem predykcji dobowego zapotrzebowania na energię z jednodniowym wyprzedzeniem czasowym:

$$ZD_d = f(ZG_{d-1}(1),...,ZG_{d-1}(24),TMIN_{d-1}, TMAX_{d-1},TMIN_d,TMAX_d,dt_{1d},...,dt_{6d}) \quad (3.3.33)$$

gdzie:

ZD_d – dobowe zapotrzebowanie na energię w dniu d,
$ZG_{d-1}(1)$, ..., $ZG_{d-1}(24)$ – godzinowy rozkład zużycia w dniu poprzednim,
$TMIN_{d-1}$, $TMAX_{d-1}$ – temperatura minimalna i maksymalna w dniu poprzednim,
$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,
dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Metodę szacowania przedziałów prognozy przetestowano dla modeli prognostycznych otrzymanych przy użyciu warstwowej sieci perceptronowej MLP oraz sieci neuronowo-rozmytej FBF. Wyniki samej predykcji dla zagadnienia (3.3.33) prezentowane były w punkcie 2.3.2 (model 1).

Tabela 3.3.2. Częstości empiryczne przedziałów prognozy dobowego zapotrzebowania na energię, otrzymanych za pomocą metody delta dla sieci MLP i FBF (w %)

Prawdopodobieństwo	Dokładny hesjan błędu		Przybliżony hesjan błędu	
	MLP	FBF	MLP	FBF
80	81,24	81,77	88,51	88,86
85	85,63	86,16	92,86	92,34
90	89,57	90,25	96,20	94,60
95	93,99	93,25	98,17	96,86

Źródło: opracowanie własne.

Podobnie jak w przypadku tabeli 3.3.1, również obecnie przedziały prognozy przetestowane zostały dla dużego zbioru niezależnych danych testowych, które nie były wykorzystywane wcześniej do budowy modelu. Dla powyższych wzorców obliczony został procent obserwacji rzeczywistego zapotrzebowania na energię w obrębie przedziałów prognozowanych dla kilku wybranych typowych poziomów prawdopodobieństwa.

Wyniki testowania znajdują się w tabeli 3.3.2. Przebadano obie wersje metody delta, dla pełnego hesjanu błędu modelu MLP oraz FBF (za pomocą zależności (3.3.30)) oraz przybliżonego hesjanu (zależność (3.3.31)). Podobnie jak w przypadku poprzednio testowanych modeli, widzimy, że oszacowania wariancji prognozy zapotrzebowania na energię oraz uzyskanych za jej pomocą przedziałów prognozy w miarę dokładnie odzwierciedlają spodziewane prawdopodobieństwa (Bartkiewicz 2011a; Bartkiewicz 2012). Zarówno dla sieci MLP, jak i FBF wyniki są wyraźnie lepsze w przypadku wykorzystania dokładnej wartości hesjanu.

Kolejne analizowane zagadnienie dotyczyć będzie prognozy zapotrzebowania na moc w szczycie wieczornym z południowym wyprzedzeniem czasowym (model 3 w punkcie 2.3.2):

$$ZSW_d = f(ZSR_{d-1}, ZSW_{d-1}, TR_{d-1}, TP_{d-1}, TW_{d-1}, ZSR_d, TR_d, TP_d, TW_d) \tag{3.3.34}$$

gdzie:

ZSW_d – zapotrzebowanie na energię w szczycie wieczornym, w dniu d,
ZSR_d – zapotrzebowanie na energię w szczycie porannym, w dniu d,
TR_d – temperatura poranna (mierzona o godzinie 8), w dniu d,
TP_d – temperatura w południe (mierzona o godzinie 13), w dniu d,
TW_d – temperatura wieczorna (mierzona o godzinie 21), w dniu d.

Tabela 3.3.3. Częstości empiryczne przedziałów prognozy mocy w szczycie wieczornym, otrzymanych za pomocą metody delta dla sieci FBF (w %)

Prawdopodobieństwo	Dokładny hesjan błędu	Przybliżony hesjan błędu
	FBF	FBF
80	81,23	90,86
85	85,83	93,62
90	89,80	95,92
95	93,64	98,49

Źródło: opracowanie własne.

W przypadku modelu (3.3.34) przetestowano przedziały prognozy otrzymane metodą delta dla sieci neuronowo-rozmytej FBF (Bartkiewicz 2012). Podobnie jak w poprzednich przypadkach, w tabeli 3.3.3 zawarto informacje o procencie obserwacji rzeczywistego zapotrzebowania na energię w obrębie przedziałów prognozy wyznaczonych dla pełnego hesjanu błędu sieci FBF (za pomocą zależności (3.3.30)) oraz przybliżonego hesjanu, oszacowanego metodą aproksymacji iloczynem skalarnym (zależność (3.3.31)).

Wyniki przedstawione w tabeli 3.3.3 ponownie potwierdzają nasze poprzednio poczynione obserwacje, że oszacowania przedziałów prognozy zapotrzebowania na energię, w których wykorzystuje się metodę delta, w miarę dokładnie odzwierciedlają spodziewane prawdopodobieństwa, przy czym wyraźnie lepsze wyniki otrzymaliśmy w przypadku zastosowania algorytmów obliczających dokładną wartość hesjanu.

Ostatni test empiryczny, jaki przeprowadziliśmy, związany jest z zastosowaniem metody delta do oszacowania wariancji wyjściowej modelu i, w konsekwencji, przedziałów prognozy dla sieci neuronowo-rozmytej typu Takagi–

Sugeno, z liniowymi następnikami reguł (Bartkiewicz 2011b; Bartkiewicz 2012). Tego typu modele wykorzystywaliśmy w punkcie 2.3.4 do prognozy szczytowego zapotrzebowania na energię (maksymalnego godzinnego zapotrzebowania na energię) z dwudniowym wyprzedzeniem czasowym:

$$ZS_d = f(ZS_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZS(t)_{d-2}, ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \tag{3.3.35}$$

gdzie oznaczenia są analogiczne jak w modelach prezentowanych poprzednio:

$ZS(t)_d$ – zapotrzebowanie na energię w szczytowej godzinie, w dniu d,

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d, gdzie t oznacza godzinę szczytową,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Zauważmy jednak, że do oszacowania wag sieci neuronowo-rozmytej typu Takagi–Sugeno, wykorzystanej w punkcie 2.3.4 do prognozy (3.3.35), nie używaliśmy bezpośrednio metody najmniejszych kwadratów. W przypadku tego modelu zastosowaliśmy dwustopniową procedurę polegającą na oszacowaniu parametrów zbiorów rozmytych poprzedników za pomocą algorytmu grupowania danych, a dopiero potem wyznaczaliśmy współczynniki funkcji liniowych następników reguł systemu metodą najmniejszych kwadratów. Oznacza to, niestety, że do określenia wariancji wyjściowej sieci, wynikającej z niepewności wag, nie możemy zastosować metody delta, która przecież oparta jest na analizie błędu kwadratowego. Metodę tę możemy stosować do sieci neuronowo-rozmytej typu Takagi–Sugeno, pod warunkiem jednak, że wszystkie jej wagi trenowane są metodą wstecznej propagacji, minimalizującą błąd kwadratowy prognozy otrzymywanej przez model. Co więcej, dla tego rodzaju modeli, z hybrydowym algorytmem uczenia parametrów, nie ma opracowanych żadnych oszacowań analitycznych wariancji wyjściowej i w teorii ograniczeni jesteśmy wyłącznie do podejść empirycznych (np. omawianych w punkcie 3.3.5 metod opartych na bootstrapie).

Postanowiliśmy jednak przetestować zastosowanie metody delta do powyższego przypadku, ograniczając się wyłącznie do części modelu uczonej metodą najmniejszych kwadratów, tj. do współczynników funkcji liniowych występujących w następnikach reguł. Postępowanie takie jest uprawnione, ponieważ współczynniki te wyznaczane są po wcześniejszym określeniu parametrów zbiorów rozmytych w poprzednikach. W trakcie finalnego uczenia modelu parametry mają charakter stały. Zaniedbujemy jednak w ten sposób wariancję prognozy wynikającą z niepewności tych parametrów.

Innymi słowy, potraktowaliśmy sieć neuronowo-rozmytą Takagi–Sugeno, podobnie zresztą jak zrobiliśmy to na potrzeby jej uczenia, jako uogólniony model regresji krzywoliniowej (2.3.28) i (2.3.29). Odwołując się do rozważań z punktu 3.2.4, dotyczących wyznaczania wariancji wyjścia dla modeli liniowych, możemy na podstawie zależności (3.2.33) napisać:

$$\sigma^2_y(\mathbf{x}) = \sigma^2_y(\mathbf{z}) = \sigma^2_\varepsilon (1 + \mathbf{z}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{z}) \tag{3.3.36}$$

gdzie, za (2.3.29), możemy zmienną $\mathbf{z}$ zdefiniować jako $(n + 1)\cdot K$-elementową zmienną wektorową:

$$\begin{aligned} z_{ij} &= v_i x_j, \quad j = 1,...,n, i = 1,...,K \\ z_{i0} &= v_i \end{aligned} \tag{3.3.37}$$

Przypomnijmy, że zgodnie z oznaczeniami w punkcie 3.2.4, K oznacza liczbę reguł, natomiast n – liczbę zmiennych wejściowych sieci neuronowo-rozmytej. Macierz $\mathbf{Z}$ w poszczególnych wierszach będzie zawierała obserwacje wzorców wejściowych $\mathbf{z}_k$ utworzonych na podstawie wejść próbek treningowych $\mathbf{x}_k$ oraz obliczonych dla nich znormalizowanych stopni prawdziwości reguł $v_i(\mathbf{x}_k)$, $k = 1, \ldots, N$:

$$\begin{aligned} \mathbf{Z} = [&v_1(\mathbf{x}_k), v_1(\mathbf{x}_k)x_{k1},...,v_1(\mathbf{x}_k)x_{kn},...,v_K(\mathbf{x}_k), v_K(\mathbf{x}_k)x_{k1},...,v_K(\mathbf{x}_k)x_{kn}], \\ &k = 1,...,N \end{aligned} \tag{3.3.38}$$

Podobnie jak w poprzednich przypadkach w bieżącym rozdziale, oszacowanie wariancji wyjściowej sieci neuronowo-rozmytej typu Takagi–Sugeno – dane przez (3.3.37) – wykorzystaliśmy do wyznaczenia i przetestowania przedziałów prognozy dla tego modelu (uwzględniając również oszacowanie wariancji czynnika losowego, zależne od wzorca wejściowego $\mathbf{x}$ – patrz dyskusja przy okazji zależności (3.3.24)):

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x}, \mathbf{w}) - Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_\varepsilon(\mathbf{x})\sqrt{1+\mathbf{z}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{z}} \\ g_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x}, \mathbf{w}) + Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_\varepsilon(\mathbf{x})\sqrt{1+\mathbf{z}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{z}} \end{aligned} \tag{3.3.39}$$

Tabela 3.3.4. Częstości empiryczne przedziałów prognozy maksymalnej wartości energii godzinnej, otrzymanych za pomocą metody delta (w %)

Prawdopodobieństwo	Częstość
80	76,95
85	83,53
90	88,25
95	92,49

Źródło: opracowanie własne.

Wyniki testowania otrzymanych przedziałów prognozy znajdują się w tabeli 3.3.4. Otrzymane częstości rzeczywistych obserwacji maksymalnej energii godzinnej w obrębie oszacowanych przedziałów prognozy okazały się całkiem poprawne. Wskazywać to może na nieco mniejsze znaczenie niepewności parametrów zbiorów rozmytych poprzedników reguł sieci neuronowo-rozmytej typu Takagi–Sugeno dla finalnej wariancji wyjścia modelu.

Podsumowując, w bieżącym punkcie przebadano zastosowanie metody delta w przypadku modeli neuronowych i neuronowo-rozmytych, wykorzystywanych w procesie krótkoterminowej prognozy zapotrzebowania na energię. Zastosowanie tego rodzaju oszacowania analitycznego wariancji prognozy wymaga zawsze weryfikacji empirycznej. Metoda ta ma bowiem charakter przybliżony; opiera się na pewnych założeniach (które w praktyce zazwyczaj spełnione są jedynie w przybliżeniu) i wykorzystuje szereg aproksymacji i uproszczeń. Pośród najważniejszych czynników, które mogą spowodować jej niewłaściwe działanie, wymienić trzeba następujące:

– metoda zakłada normalny charakter rozkładu błędu losowego, a co za tym idzie rozkładu prognozy,

– zakładamy, że model jest nieobciążony i poprawnie zbudowany,

– metoda opiera się na lokalnej linearyzacji nieliniowego modelu neuronowego lub neuronowo-rozmytego,

– wykorzystujemy nie do końca uprawnione zastąpienie stałej wariancji rozkładu błędu losowego wariancją zmienną, zależną od wartości wzorca wejściowego danej prognozy,

– metodę tę możemy wykorzystać tylko, gdy dla danego modelu potrafimy oszacować hesjan błędu.

Metodę delta wykorzystaliśmy do oszacowania wariancji wyjściowej prognozy, a następnie do uzyskania przedziałów prognozy dla szeregu różnych zadań prognostycznych z obszaru krótkoterminowego prognozowania obciążeń sieci oraz dla kilku modeli neuronowych i neuronowo-rozmytych wykorzystywanych do ich rozwiązywania. Pomimo przedstawionej listy zastrzeżeń,

uzyskane wyniki empiryczne wskazują, że metoda delta dla rozważanych w pracy zagadnień może dawać oszacowania rozkładu prognozy zbliżone do rzeczywistych. Warunek kluczowy stanowi tu, oczywiście, założenie pierwsze, dotyczące normalnego charakteru rozkładu prawdopodobieństwa prognozy, które było spełnione we wszystkich badanych przypadkach.

Przeprowadzone badania wskazują również na to, że wykorzystanie w metodzie delta przybliżonej macierzy hesjanu błędu modelu, uzyskanej metodą aproksymacji iloczynem skalarnym (Levenberga–Marquarda), ma wyraźny wpływ na pogorszenie dokładności wyznaczonej wartości wariancji wyjściowej prognozy. Wyniki otrzymane dla przypadku przybliżonego nie są może zupełnie odmienne od rzeczywistości, ale dokładność otrzymanych oszacowań okazała się wyraźnie gorsza. Często podnosi się tu jako argument fakt, że nakłady obliczeniowe wymagane do pełnego obliczenia hesjanu są znacznie wyższe niż w przypadku aproksymacji iloczynem skalarnym. Należy jednak zauważyć, że biorąc pod uwagę raczej niewielkie rozmiary wykorzystywanych w prognozach zapotrzebowania na energię architektur sieci neuronowych i neuronowo-rozmytych oraz moce obliczeniowe współczesnych komputerów, argument ten przestaje mieć praktyczne znaczenie.

Czasy realizacji algorytmów metody delta w przypadku oszacowań pełnego hesjanu błędu zamykały się w ułamkach sekundy, pomimo wykorzystywania do odwracania macierzy hesjanu odpornego, ale wolniejszego algorytmu rozkładu na wartości osobliwe (SVD). Nie zaobserwowano również problemów ze stabilnością odwracania macierzy w omawianych przypadkach. Ponadto pamiętać należy, że macierz kowariancji parametrów wyznacza się raz, po zakończeniu treningu modelu (czas realizacji tej części obliczeń jest niemal nieznaczący w porównaniu z czasem treningu sieci neuronowej czy neuronowo-rozmytej metodą minimalizacji błędu kwadratowego). Proces ten nie musi być powtarzany przy każdej prognozie; jeśli nie ma konieczności ponownego uczenia modelu, wykorzystujemy tę samą macierz kowariancji.

Biorąc pod uwagę przedstawione czynniki, zdecydowanie sugerujemy wykorzystanie w metodzie delta algorytmów obliczeń pełnego hesjanu błędu modelu, co w badanych przypadkach pozwalało na wyznaczenie poprawnych, zbliżonych do rzeczywistych, wartości odchylenia standardowego, a, co za tym idzie, rozkładu prawdopodobieństwa prognozy. Jeszcze raz jednak należy zwrócić uwagę na to, że przeprowadzone badania stanowią poważną wskazówkę na rzecz możliwości użycia metody delta w zadaniach krótkoterminowego prognozowania zapotrzebowania na energię – jej zastosowanie wymaga jednak zawsze przeprowadzenia weryfikacji empirycznej dla każdego konkretnego przypadku.

3.3.3. Oszacowanie kanapkowe

Tak zwany estymator kanapkowy (*sandwich estimator*) (Efron, Tibshirani 1993; White 1994; Tibshirani 1996), nazywany również czasami „estymatorem kanapkowym Hubera", stanowi jedną z najbardziej znanych metod określania wariancji oszacowań rozmaitych wielkości wyznaczanych metodą największej wiarygodności w sytuacji, w której model nie jest w pełni poprawny. Podejście to zalicza się do grupy tzw. odpornych metod estymacji i wykorzystuje się w nim częściowo informacje empiryczne. Dzięki temu estymator kanapkowy może dawać lepsze oceny wariancji (kowariancji) znalezionych parametrów modelu nie tylko wtedy, gdy nie do końca spełnione są założenia dotyczące charakteru samego modelu, ale również w sytuacji, w której pojawiają się problemy związane z rozkładem prawdopodobieństwa tychże parametrów. Zaleca się go m.in. stosować w przypadku zmiennej wariancji (heteroskedastyczności) modelu (White 1994).

Nie będziemy tu oczywiście szczegółowo prezentować całej teorii matematycznej oszacowania kanapkowego, przedstawimy jedynie szkic wyprowadzenia metody, który daje jej ogólne uzasadnienie. Podobnie jak w poprzednim punkcie 3.3.2, dla metody delta, wychodzimy od aproksymacji kwadratowej błędu modelu (3.3.1) na zbiorze treningowym za pomocą rozwinięcia w szereg Taylora, w pobliżu zestawu „prawdziwych" wag modelu $\mathbf{w}^*$ (zależność (3.3.7)):

$$E(\mathbf{w}) = E(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^{\mathrm{T}} \nabla E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^{\mathrm{T}} \nabla\nabla E(\mathbf{w}^*)(\mathbf{w}-\mathbf{w}^*) \qquad (3.3.40)$$

Jeżeli model nie jest w pełni poprawny, to nie znajdziemy „prawdziwych" wag modelu $\mathbf{w}^*$, lecz pewien zestaw wag $\mathbf{w}$ minimalizujących błąd (3.3.40). Możemy go otrzymać, przyrównując pochodną błędu do 0:

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \nabla E(\mathbf{w}^*) + \nabla\nabla E(\mathbf{w}^*)(\mathbf{w}-\mathbf{w}^*) = 0 \qquad (3.3.41)$$

Obliczając teraz z (3.3.41) $\mathbf{w} - \mathbf{w}^*$, otrzymujemy:

$$\mathbf{w} - \mathbf{w}^* = -\nabla\nabla E(\mathbf{w}^*)^{-1} \nabla E(\mathbf{w}^*) \qquad (3.3.42)$$

Zmiana wyznaczonych w procesie uczenia wag $\mathbf{w}$ wokół ich „prawdziwej" wartości $\mathbf{w}^*$ jest więc proporcjonalna do wartości gradientu błędu treningowego sieci w przestrzeni wag, przetransformowanej za pomocą przekształcenia liniowego, którego macierz dana jest przez odwrotność hesjanu błędu $-\nabla\nabla E(\mathbf{w}^*)^{-1}$. Oznaczając tradycyjnie hesjan błędu treningowego względem wag

przez **H** oraz korzystając z prawa propagacji błędów (3.2.29), które wprowadziliśmy w punkcie 3.2.4, otrzymujemy następującą zależność dla macierzy kowariancji błędów modelu:

$$\begin{aligned} \mathbf{C}_{\mathbf{w}} &= (-\mathbf{H}^{-1})\,\mathrm{cov}(\nabla E(\mathbf{w}^*))(-\mathbf{H}^{-1}) \\ \mathbf{C}_{\mathbf{w}} &= \mathbf{H}^{-1}\,\mathrm{cov}(\nabla E(\mathbf{w}^*))\mathbf{H}^{-1} \end{aligned} \tag{3.3.43}$$

Zależność (3.3.44) wyjaśnia przy tym, skąd się wzięła nazwa oszacowanie (estymator) kanapkowe (*sandwich estimator*). Odwrotność hesjanu otacza w nim z obu stron macierz kowariancji gradientów błędów treningowych względem wag, $\mathrm{cov}(\nabla E(\mathbf{w}^*))$, co przypomina klasyczny sandwicz, obłożony chlebem. Jak widać, poczucie humoru niektórych naukowców powinno chyba stać się przedmiotem bliższego zainteresowania profesjonalistów z odpowiedniej dziedziny.

Macierz kowariancji gradientów błędów $\mathrm{cov}(\nabla E(\mathbf{w}^*))$ wyznacza się empirycznie na podstawie gradientów błędów w próbie (zbiorze uczącym sieci). Jeżeli więc zbiór $D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$, jest zbiorem treningowym wykorzystanym do dopasowania modelu, to gradient błędu kwadratowego dla dowolnego k-tego wzorca uczącego $\nabla E(\mathbf{w}^*, \mathbf{x}_k)$ wyznaczyć można za pomocą zależności:

$$\nabla E(\mathbf{w}^*, \mathbf{x}_k) = \mathbf{s}_k \tag{3.3.44}$$

gdzie $\mathbf{s}_k = (s_{k1}, \ldots, s_{kp})^{\mathrm{T}}$ jest wektorem pochodnych cząstkowych błędu względem poszczególnych parametrów (wag) modelu:

$$\begin{aligned} s_{ki} &= \left.\frac{\partial E(\mathbf{w}, \mathbf{x}_k)}{\partial w_i}\right|_{\mathbf{w}=\mathbf{w}^*} = \left.\frac{\partial}{\partial w_i}\left(\frac{1}{2}(y_k - f(\mathbf{x}_k, \mathbf{w}))^2\right)\right|_{\mathbf{w}=\mathbf{w}^*} = \\ &= -(y_k - f(\mathbf{x}_k, \mathbf{w}))\left.\frac{\partial f(\mathbf{x}_k, \mathbf{w})}{\partial w_i}\right|_{\mathbf{w}=\mathbf{w}^*} \qquad ,i = 1,\ldots,p \end{aligned} \tag{3.3.45}$$

zaś p jest liczbą parametrów.

Macierz kowariancji gradientów błędów $\mathrm{cov}(\nabla E(\mathbf{w}^*))$ możemy wtedy oszacować przy użyciu empirycznej kowariancji gradientów błędów:

$$\mathrm{cov}(\nabla E(\mathbf{w}^*)) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{s}_k\mathbf{s}_k^T \tag{3.3.46}$$

Zauważmy przy tym, że kowariancje (3.3.46) nie muszą być centrowane, ponieważ wartość oczekiwana gradientu błędu w punkcie jego minimum jest równa zero.

Ostatecznie więc kanapkowe oszacowanie macierzy kowariancji wag modelu $\mathbf{C}_{\mathbf{w}}$, które oznaczać będziemy $\mathbf{V}_{sand}$, możemy na podstawie (3.3.43) i (3.3.46) zapisać za pomocą następującej zależności:

$$\mathbf{V}_{sand} = \mathbf{H}^{-1}\left(\frac{1}{N}\sum_{k=1}^{N}\mathbf{s}_k\mathbf{s}_k^T\right)\mathbf{H}^{-1} \tag{3.3.47}$$

gdzie $\mathbf{s}_k$, k =1, …, N dane są przez gradienty błędów modelu dla poszczególnych wzorców treningowych (3.3.45).

Otrzymaliśmy więc oszacowanie macierzy kowariancji parametrów modelu w postaci estymatora kanapkowego $\mathbf{V}_{sand}$. Aby wyznaczyć wariancję prognozy wynikającą z niepewności parametrów, stosujemy dalej identyczne podejście jak w przypadku analizowanej w poprzednim punkcie metody delta, oparte na lokalnej linearyzacji modelu. Wykorzystujemy więc bezpośrednio zależność (3.3.5):

$$\begin{aligned}\sigma_{\mathbf{w}}^2(\mathbf{x}) &= \mathbf{g}(\mathbf{x},\mathbf{w}^*)^{\mathrm{T}}\mathbf{C}_w\mathbf{g}(\mathbf{x},\mathbf{w}^*) = \mathbf{g}(\mathbf{x},\mathbf{w}^*)^{\mathrm{T}}\mathbf{V}_{sand}\mathbf{g}(\mathbf{x},\mathbf{w}^*) = \\ &= \mathbf{g}(\mathbf{x},\mathbf{w}^*)^{\mathrm{T}}\mathbf{H}^{-1}\left(\frac{1}{N}\sum_{k=1}^{N}\mathbf{s}_k\mathbf{s}_k^T\right)\mathbf{H}^{-1}\mathbf{g}(\mathbf{x},\mathbf{w}^*)\end{aligned} \tag{3.3.48}$$

Estymator kanapkowy macierzy kowariancji wag przetestowaliśmy dla zagadnień krótkoterminowego prognozowania zapotrzebowania na energię elektryczną, z wykorzystaniem predyktora w formie sieci neuronowej MLP i neuronowo-rozmytej FBF. Podobnie jak w poprzednich przypadkach, przebadaliśmy dokładność oszacowań wariancji wyjściowej modelu na przykładzie przedziałów prognozy.

$$\begin{aligned}d_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) - Q_{N(0,1)}((1+\alpha)/2)\cdot\sqrt{\sigma_\varepsilon^2(\mathbf{x}) + \sigma_{\mathbf{w}}^2(\mathbf{x})} \\ g_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) + Q_{N(0,1)}((1+\alpha)/2)\cdot\sqrt{\sigma_\varepsilon^2(\mathbf{x}) + \sigma_{\mathbf{w}}^2(\mathbf{x})}\end{aligned} \tag{3.3.49a}$$

gdzie wariancja prognozy wynikająca z niepewności wag $\sigma^2_{\mathbf{w}}(\mathbf{x})$ określona jest za pomocą zależności (3.3.48). Odchylenie standardowe elementu losowego $\sigma_\varepsilon(\mathbf{x})$, dla danego wzorca wejściowego prognozy $\mathbf{x}$, oszacowane zostało z wykorzystaniem dodatkowego modelu, zgodnie z dyskusją przedstawioną w kolejnym podrozdziale 3.4. Wartość $Q_{N(0,1)}(\alpha)$ oznacza, podobnie jak w poprzednich

przypadkach, α-kwantyl standardowego normalnego rozkładu prawdopodobieństwa $N(0, 1)$. Ponownie przypomnijmy, że z powodu dużych rozmiarów zbioru treningowego w oszacowaniu przedziałów prognozy mogliśmy wykorzystać kwantyl rozkładu normalnego prognozy. W przypadku mniejszej liczby wzorców treningowych należałoby wykorzystać kwantyl rozkładu t-Studenta.

Ponadto, zamiast zależności (3.3.49a), w której normalizujemy otrzymany rozkład prognozy i wykorzystujemy do wyznaczenia krańców jej przedziałów kwantyle standardowego rozkładu normalnego prawdopodobieństwa $Q_{N(0,1)}(\alpha)$, możemy bezpośrednio zastosować rozkład prawdopodobieństwa prognozy $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$. Krańce przedziału otrzymujemy, obliczając odpowiednie kwantyle rozkładu:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= Q_{N(f(\mathbf{x},\mathbf{w}),\sigma_y(\mathbf{x}))}((1-\alpha)/2) \\ g_{y/\mathbf{x}}(\alpha) &= Q_{N(f(\mathbf{x},\mathbf{w}),\sigma_y(\mathbf{x}))}((1+\alpha)/2) \end{aligned} \tag{3.3.49b}$$

Przedziały prognozy określono dla przewidywanego dobowego zapotrzebowania na energię z jednodniowym wyprzedzeniem czasowym (dla tego samego zagadnienia, co model (2.3.33) w punkcie 3.3.2):

$$\begin{aligned} ZD_d = f(&ZG_{d-1}(1),...,ZG_{d-1}(24), TMIN_{d-1}, \\ &TMAX_{d-1}, TMIN_d, TMAX_d, dt_{1d},...,dt_{6d}) \end{aligned} \tag{3.3.50}$$

gdzie:

ZD_d – dobowe zapotrzebowanie na energię w dniu d,
$ZG_{d-1}(1)$, ..., $ZG_{d-1}(24)$ – godzinowy rozkład zużycia w dniu poprzednim,
$TMIN_{d-1}$, $TMAX_{d-1}$ – temperatura minimalna i maksymalna w dniu poprzednim,
$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,
dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Wyniki testowania empirycznych częstości przedziałów prognozy otrzymanych za pomocą zależności (3.3.49) dla sieci neuronowej MLP i neuronowo-rozmytej FBF przedstawione zostały w tabeli 3.3.5. Zawiera ona porównanie względnej liczby rzeczywistych obserwacji dobowego zapotrzebowania na energię w obrębie prognozowanych przedziałów dla oszacowania kanapkowego macierzy kowariancji parametrów modelu oraz macierzy kowariancji otrzymanej za pomocą metody delta, przy dokładnym oszacowaniu hesjanu błędu (dane przeniesione z tabeli 3.3.2). Zauważmy, że wyznaczenie hesjanu błędu niezbędne jest również w przypadku estymatora kanapkowego. Stanowi on przecież pewną korektę empiryczną oszacowania macierzy kowariancji opartej na hesjanie błędu. Również i tutaj wykorzystano algorytmy obliczania dokładnej jego wartości.

Tabela 3.3.5. Częstości empiryczne przedziałów prognozy dobowego zapotrzebowania na energię, otrzymanych za pomocą metody delta i estymatora kanapkowego, dla sieci MLP i FBF (w %)

Prawdopodobieństwo	Metoda delta		Oszacowanie kanapkowe	
	MLP	FBF	MLP	FBF
80	81,24	81,77	85,50	76,43
85	85,63	86,16	88,81	83,91
90	89,57	90,25	92,32	88,28
95	93,99	93,25	95,33	91,74

Źródło: opracowanie własne.

Analizując wyniki przedstawione w tabeli 3.3.5, można zauważyć, że korekta macierzy kowariancji w formie oszacowania kanapkowego nie do końca zdała egzamin w testowanych przypadkach. Metoda uwzględniania heteroskedastyczności modelu poprzez dostarczenie dodatkowego oszacowania odchylenia standardowego błędu losowego prognozy zależnego od jej wejścia, jaką zastosowaliśmy w metodzie delta, działa wyraźnie lepiej niż korekta tego zjawiska poprzez oszacowanie kanapkowe. Należy jednak podkreślić, że wyniki doświadczalne w badanych przypadkach nie odbiegają na tyle od oczekiwanych wartości, by całkowicie negować przydatność estymatora kanapkowego jako narzędzia wyznaczania wariancji rozkładu prawdopodobieństwa zapotrzebowania na energię. Wydaje się, że jest to metoda mniej dokładna, ale szybsza, w tym sensie, że nie wymaga niezbędnie znalezienia modelu dosyć złożonego odchylenia błędu losowego, co w przypadku metody delta powoduje konieczność tworzenia dodatkowego modelu prognostycznego (patrz podrozdział 3.4).

3.3.4. Oszacowanie wariancji prognozy z wykorzystaniem bootstrapu

Poprzednio omawiane podejścia do określania niepewności prognozy zapotrzebowania na energię miały charakter albo w pełni analityczny, jak metoda delta, albo mieszany (w oszacowaniu kanapkowym). Podejście oparte na bootstrapie, dla odmiany, stanowi rozwiązanie całkowicie oparte na ocenach empirycznych i pozwala na określenie wariancji (odchylenia standardowego) prognozy w zasadzie wyłącznie na drodze doświadczalnej, poprzez odpowiednie techniki wielokrotnego tworzenia i testowania modelu. Znika przy tym, rzecz jasna, większość założeń czynionych dla poprzednio prezentowanych metod. Jedynym, które nadal pozostaje w mocy, jest założenie o normalnym charakterze rozkładu prognozy. Kosztem związanym ze stosowaniem bootstrapu są znacznie większe nakłady obliczeniowe niż te, które ponosiliśmy w przypadku metod analitycznych.

Jeżeli chcemy oprzeć się wyłącznie na badaniach empirycznych, to nietrudno się domyślić, że największy problem będziemy mieli z oszacowaniem niepewności wag sieci neuronowej czy też neuronowo-rozmytej. Niepewność ta wynika bowiem przede wszystkim ze skończonego charakteru próby losowej, zbioru treningowego, na którym uczymy tworzone modele, w konfrontacji z nieskończoną populacją ogólną. W związku z tym, testując zbudowany model, otrzymujemy informacje na temat działania i posiadanych właściwości wyłącznie dla tego konkretnego modelu, dostosowanego do tej konkretnej serii danych, na której został stworzony. W poprzednio przedstawianych metodach analizujemy ten zbiór danych w konfrontacji ze strukturą i charakterystyką modelu, próbując na tej podstawie przeprowadzić wnioskowanie o jego zachowaniu dla wzorców, które są podobne do treningowych, ale nie są dokładnie takie same. Jeżeli chcemy tego uniknąć – w zasadzie nie ma innego wyjścia – nie możemy ograniczyć się do badania jednego modelu i jednej próby.

Właśnie bootstrap jest znaną i w sumie dosyć prostą (przynajmniej jeżeli chodzi o stosowanie) techniką statystyczną ogólnego przeznaczenia, która służy do redukowania, niwelowania, szacowania itp. efektów skończonego charakteru próby wykorzystywanej przy wyznaczaniu różnego rodzaju oszacowań statystycznych (Efron, Tibshirani 1993). Mówiąc ogólnie, bootstrap polega na wygenerowaniu nie jednej, ale wielu prób oraz, w konsekwencji, uzyskaniu za ich pomocą wielu estymatorów szukanej wielkości, a następnie scaleniu ich w jedno ostateczne oszacowanie. W związku z tym technika ta stanowi przykład podejścia opartego na wielokrotnym powtarzaniu próbkowania populacji ogólnej (*resampling*). Przydatność takiego sposobu postępowania do uchwycenia zmienności modelu oszacowania i samego estymatora szukanej wielkości, w zależności od różnic w próbie użytej do ich otrzymania, wydaje się raczej dosyć jasna i oczywista i nie wymaga żadnych dodatkowych komentarzy.

Pamiętajmy jednak, że do celów budowy modelu prognostycznego zakładamy posiadanie jednej próby, jednego zbioru danych treningowych $D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$, który ma służyć do dopasowania sieci neuronowej czy też neuronowo-rozmytej. Powstaje wobec tego pytanie: w jaki sposób możemy zastosować bootstrap do analizy tego typu modeli. Rozwiązanie polega na wielokrotnym próbkowaniu (losowaniu z powtórzeniami wzorców) zbioru D, a następnie wykorzystaniu go jako podstawy do wygenerowania szeregu serii danych, na których nauczona zostanie cała grupa sieci.

Efron i Tibshirani wyróżniają dwa podstawowe podejścia do wykorzystania tej techniki na potrzeby zadań regresyjnych (Efron, Tibshirani 1993; Tibshirani 1996). Pierwsze z nich, określane jako „bootstrapowanie poprzez próbkowanie par", polega zasadniczo na tworzeniu prób uczących dla kolejnych modeli bezpośrednio poprzez wylosowanie ich z powtórzeniami z wyjściowego zbioru danych treningowych D. Drugie podejście, tzw. „bootstrapowanie poprzez próbkowanie reszt", polega na losowaniu prób dla kolejnych modeli ze zbioru

reszt pewnego modelu bazowego, otrzymanych dla wzorców treningowych ze zbioru D. Bardziej szczegółowo mówiąc, obydwa sugerowane algorytmy zdefiniować można w przedstawiony dalej sposób (za: Tibshirani 1996).

Bootstrapowanie poprzez próbkowanie par

1. Próbkujemy wyjściowy zbiór treningowy, to znaczy z posiadanego zbioru danych treningowych $D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$ wybieramy losowo ze zwracaniem (z jednostajnym rozkładem prawdopodobieństwa) B prób D_i, z których każda składa się z N wzorców uczących.

2. Dla każdej z otrzymanych wcześniej prób D_i trenujemy odrębną sieć neuronową lub neuronowo-rozmytą $f(\mathbf{x}, \mathbf{w}_i)$, $i = 1, \ldots, K$.

3. Dla dowolnego wzorca wejściowego $\mathbf{x}$, wariancję wyjścia sieci (otrzymywanej prognozy), spowodowaną niepewnością oszacowanego dla modelu zestawu parametrów $\sigma_{\mathbf{w}}^2(\mathbf{x})$, możemy wyznaczyć jako:

$$\sigma_{\mathbf{w}}^2(\mathbf{x}) = \frac{1}{B-1}\sum_{i=1}^{B}(f(\mathbf{x}, \mathbf{w}_i) - f_{sr}(\mathbf{x}, \cdot))^2 \tag{3.3.51}$$

gdzie $f_{sr}(\mathbf{x},)$ jest średnią prognoz otrzymywanych ze wszystkich bootstrapowanych modeli $f(\mathbf{x}, \mathbf{w}_i)$, $i = 1, \ldots, K$:

$$f_{sr}(\mathbf{x}, \cdot) = \frac{1}{B}\sum_{i=1}^{B} f(\mathbf{x}, \mathbf{w}_i) \tag{3.3.52}$$

Jako oszacowanie wartości oczekiwanej działania modelu wykorzystujemy we wzorze (3.3.52) średnią arytmetyczną z wyjść sieci otrzymanych dla poszczególnych bootstrapowanych zbiorów treningowych. Podejście to określa się zazwyczaj jako „pakowanie" (*bagging*). W niektórych przypadkach lepsze wyniki daje stosowanie różnego rodzaju średnich ważonych.

Jak widzimy bootstrapowanie poprzez próbkowanie par odbywa się w zasadzie zupełnie w oderwaniu od jakiegokolwiek wyjściowego modelu prognostycznego. Oszacowanie wariancji otrzymywanej prognozy, wynikającej z niepewności parametrów modelu, odbywa się wyłącznie na podstawie cech charakterystycznych zbioru danych treningowych wykorzystanych do jego budowy. Z odmienną sytuacją mamy do czynienia w przypadku podejścia opartego na bootstrapowaniu reszt.

Bootstrapowanie poprzez próbkowanie reszt

1. Za pomocą zbioru treningowego $D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$, tworzymy nasz model prognostyczny, dopasowując do danych sieć neuronową lub neuronowo-rozmytą $f(\mathbf{x}, \mathbf{w})$.

2. Dla gotowej sieci $f(\mathbf{x}, \mathbf{w})$ obliczamy reszty modelu dla wszystkich wzorców treningowych:

$$e_k = y_k - f(\mathbf{x}_k, \mathbf{w}),\ k = 1, \ldots, N$$

3. Próbkujemy zbiór otrzymanych reszt modelu prognostycznego, to znaczy ze zbioru wszystkich reszt losujemy ze zwracaniem (przy jednostajnym rozkładzie prawdopodobieństwa) B prób, z których każda składa się z N elementów. Oznaczmy dowolną i-tą próbę przez $e^i_1, e^i_2, \ldots, e^i_N$, $i = 1, \ldots, B$.

4. Na podstawie treningowych wzorców wejściowych $\mathbf{x}_k$ ze zbioru D dla poszczególnych bootstrapowanych prób tworzymy zbiory treningowe D_i, przyjmując jako pożądane wyjścia treningowe w procesie uczenia prognozy otrzymywane z modelu, zaburzone wylosowanymi błędami resztowanymi:

$$D_i = \{\mathbf{x}_k, f(\mathbf{x}_k, \mathbf{w}) + e^i_k\},\ k = 1, \ldots, N, \quad i = 1, \ldots, B$$

5. W kolejnych krokach algorytmu postępujemy już w zasadzie tak samo jak w przypadku bootstrapowania poprzez próbkowanie par. Dla każdej z otrzymanych wcześniej prób D_i tworzymy więc odrębny model, ucząc na niej sieć neuronową lub neuronowo-rozmytą $f(\mathbf{x}, \mathbf{w}_i)$, $i = 1, \ldots, K$.

6. Dla dowolnego wzorca wejściowego $\mathbf{x}$, wariancję wyjścia sieci (otrzymywanej prognozy), spowodowaną niepewnością oszacowanego dla modelu zestawu parametrów $\sigma_{\mathbf{w}}^2(\mathbf{x})$, możemy wyznaczyć jako:

$$\sigma^2_{\mathbf{w}}(\mathbf{x}) = \frac{1}{B-1}\sum_{i=1}^{B}(f(\mathbf{x}, \mathbf{w}_i) - f_{sr}(\mathbf{x}, \cdot))^2 \tag{3.3.53}$$

gdzie $f_{sr}(\mathbf{x},)$ jest średnią prognoz otrzymywanych ze wszystkich bootstrapowanych modeli $f(\mathbf{x}, \mathbf{w}_i)$, $i = 1, \ldots, K$:

$$f_{sr}(\mathbf{x}, \cdot) = \frac{1}{B}\sum_{i=1}^{B} f(\mathbf{x}, \mathbf{w}_i) \tag{3.3.54}$$

Jak można zauważyć, w przypadku bootstrapowania poprzez próbkowanie reszt oszacowanie wariancji prognozy odbywa się ściśle na bazie modelu. Która z zaprezentowanych metod jest lepsza? Nie ma tutaj jednoznacznej odpowiedzi. Procedura bootstrapowania reszt, jako oparta na modelu, polega na założeniu, że reszty są reprezentatywne dla prawdziwych błędów modelu. Jeżeli model nie jest w pełni poprawny, podejście oparte na bootstrapowaniu par może być mniej czułe na naruszenia jego założeń. Z drugiej jednak strony, w niektórych sytuacjach wykorzystywanie w procesie wnioskowania na temat błędu prognoz

opartych na różnych predyktorach budowanych z bootstrapowanych prób, zamiast danego z góry zbioru prognoz z jednego ustalonego modelu, może być nieodpowiednie. Problemy tego typu rodzić się będą jednak raczej w różnego rodzaju zadaniach statystycznych związanych z budową eksperymentu. W zagadnieniach prognostycznych nie powinny mieć one większego praktycznego znaczenia (Tibshirani 1996).

Analizując zastosowanie metody empirycznej opartej na bootstrapie do szacowania wariancji (odchylenia standardowego) prognozy krótkoterminowego zapotrzebowania na energię elektryczną, zdecydowaliśmy się na zastosowanie podejścia, w którym wykorzystuje się próbkowanie par. Podobnie jak w poprzednich punktach bieżącego podrozdziału, do weryfikacji działania metody wybraliśmy zagadnienie wyznaczania przedziałów prognozy dla szeregu omawianych w rozdziale 2 zadań prognostycznych oraz modeli neuronowych i neuronowo-rozmytych stosowanych do ich rozwiązania.

W przypadku wyników badań omawianych w bieżącym punkcie dla każdej prognozy zapotrzebowania, określonej przez wyjście sieci neuronowej (neuronowo-rozmytej) $f(\mathbf{x}, \mathbf{w})$, jej przedziały dla zadanego poziomu prawdopodobieństwa α wyznaczone zostały zgodnie ze wzorem:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) - Q_{N(0,1)}((1+\alpha)/2)\cdot\sqrt{\sigma_{\varepsilon}^2(\mathbf{x})+\sigma_{\mathbf{w}}^2(\mathbf{x})} \\ g_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) + Q_{N(0,1)}((1+\alpha)/2)\cdot\sqrt{\sigma_{\varepsilon}^2(\mathbf{x})+\sigma_{\mathbf{w}}^2(\mathbf{x})} \end{aligned} \tag{3.3.55a}$$

gdzie wariancja prognozy wynikająca z niepewności wag $\sigma^2_{\mathbf{w}}(\mathbf{x})$ określona została za pomocą metody bootstrapowania poprzez próbkowanie par, czyli pierwszego z prezentowanych algorytmów i zależności (3.3.51). We wszystkich opisywanych w bieżącym podrozdziale przypadkach liczba bootstrapowanych prób wykorzystywanych do oszacowania $\sigma^2_{\mathbf{w}}(\mathbf{x})$ wynosiła 30 ($B = 30$). Wariancja czynnika losowego $\sigma^2_{\varepsilon}(\mathbf{x})$ dla danego wzorca wejściowego prognozy $\mathbf{x}$, a dokładniej mówiąc, odchylenie standardowe elementu losowego $\sigma_{\varepsilon}(\mathbf{x})$, oszacowane zostało z wykorzystaniem dodatkowego estymatora, zgodnie z dyskusją przedstawioną w kolejnym podrozdziale 3.4. Wartość $Q_{N(0,1)}(\alpha)$ oznacza, tak samo jak we wszystkich poprzednio prezentowanych przypadkach, α-kwantyl standardowego normalnego rozkładu prawdopodobieństwa $N(0, 1)$.

Jak zwykle zwróćmy uwagę, że z powodu dużych rozmiarów zbioru treningowego w oszacowaniu przedziałów prognozy mogliśmy wykorzystać kwantyl rozkładu normalnego prawdopodobieństwa $Q_{N(0,1)}(\alpha)$. W przypadku mniejszej liczby wzorców treningowych należałoby zastosować kwantyl rozkładu t-Studenta. Ponadto zamiast zależności (3.3.55a), które normalizują rozkład prognozy i w których do wyznaczenia jej przedziałów korzysta się ze standardowego rozkładu normalnego prawdopodobieństwa $N(0, 1)$, możemy bezpośrednio zastosować rozkład prawdopodobieństwa prognozy $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$,

gdzie wariancja jest równa $\sigma^2_y(\mathbf{x}) = \sigma^2_{\mathbf{w}}(\mathbf{x}) + \sigma^2_{\varepsilon}(\mathbf{x})$. Krańce przedziału otrzymujemy, obliczając odpowiednie kwantyle rozkładu:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= Q_{N\left(f(\mathbf{x},\mathbf{w}),\sigma_y(\mathbf{x})\right)}((1-\alpha)/2) \\ g_{y/\mathbf{x}}(\alpha) &= Q_{N\left(f(\mathbf{x},\mathbf{w}),\sigma_y(\mathbf{x})\right)}((1+\alpha)/2) \end{aligned} \tag{3.3.55b}$$

Jako pierwsze zagadnienie praktyczne z zakresu problematyki krótkoterminowej prognozy zapotrzebowania na energię, dla którego zweryfikujemy zastosowanie bootstrapu do wyznaczenia wariancji (odchylenia standardowego) warunkowego rozkładu prawdopodobieństwa wyjścia modelu prognostycznego dla danego wzorca wejściowego, rozważmy dyskutowane już wcześniej w poprzednim rozdziale, w punktach 2.2.4 i 2.2.5, zadanie prognozy godzinnego zapotrzebowania na energię elektryczną z dwudniowym wyprzedzeniem czasowym:

$$\begin{aligned} ZG_d(t) = f(&ZG(t)_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZG(t)_{d-2}, \\ &ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \end{aligned} \tag{3.3.56}$$

gdzie oznaczenia są takie same jak w punkcie 2.2.4:

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d,
$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,
dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Przypomnijmy, że do prognozy zapotrzebowania na energię (3.3.56) wykorzystywaliśmy oddzielne sztuczne sieci neuronowe MLP (warstwowe sieci perceptronowe) o strukturze {13, 10, 1} (13 neuronów w warstwie wejściowej, 10 w warstwie ukrytej, 1 neuron wyjściowy). Dla tych modeli porównaliśmy częstości empiryczne przedziałów prognozy wyznaczonych za pomocą metody delta (dla pełnego hesjanu błędu za pomocą zależności (3.3.30)) oraz uzyskane z wykorzystaniem bootstrapu (zależność (3.3.55)) (Bartkiewicz 2001a; Bartkiewicz, Gontar, Matusiak, Zieliński 2002). Wyniki testowania dla metody delta prezentowane już były wcześniej w punkcie 3.3.2 poświęconym tej metodzie (tabela 3.3.1).

Porównanie wyników testowania obu metod znajduje się w tabeli 3.3.6. Jak już wcześniej nadmienialiśmy w punkcie 3.3.2, w procesie tym wykorzystano długi, niemal trzyletni, zbiór danych. Dane te obejmowały wszystkie rodzaje dni tygodnia, włączając w to również soboty i niedziele, odrzucono jednakże święta narodowe i religijne (oraz dni występujące bezpośrednio po nich). Tego rodzaju dni nietypowe prognozowane były przy użyciu specjalnego podejścia, przez odrębne modele (dokładniejsze wyjaśnienia znaleźć można w punkcie 2.2.5).

Analizując dane znajdujące się w tabeli 3.3.6, stwierdzamy, że wyniki uzyskane za pomocą obu analizowanych podejść mają charakter porównywalny. Można jednak zauważyć, że częstości przedziałów prognozy godzinnego zapotrzebowania na energię elektryczną uzyskane przy użyciu metody delta w niektórych przypadkach kształtują się nieco bliżej spodziewanych teoretycznych poziomów prawdopodobieństwa, dla których je wyznaczono, niż częstości przedziałów uzyskanych za pomocą bootstrapu. Potwierdza to wysnuty w punkcie 3.3.2 wniosek, że założenia przyjmowane dla metody analitycznej w badanych przypadkach zostały spełnione w stopniu dostatecznym i doprowadziły do uzyskania w miarę poprawnych wyników.

Kolejnym analizowanym zagadnieniem, dla którego przetestowaliśmy przedziały prognozy otrzymane metodą empiryczną, za pomocą bootstrapu, był problem predykcji dobowego zapotrzebowania na energię z jednodniowym wyprzedzeniem czasowym.

Tabela 3.3.6. Częstości empiryczne przedziałów prognozy godzinnego zapotrzebowania na energię, otrzymanych dla sieci MLP za pomocą metody delta i bootstrapu (w %)

Prawdopodobieństwo	Metoda delta				Bootstrap			
	godz. 7	godz. 12	godz. 17	godz. 20	godz. 7	godz. 12	godz. 17	godz. 20
80	78,21	82,32	78,45	80,87	82,80	85,20	86,10	82,20
85	84,02	85,84	82,93	84,02	87,20	89,90	89,90	83,70
90	88,38	89,23	87,77	88,86	92,00	92,90	93,50	88,70
95	93,46	93,10	92,37	92,74	95,80	96,40	96,10	94,40

Źródło: opracowanie własne.

$$ZD_d = f(ZG_{d-1}(1),...,ZG_{d-1}(24),TMIN_{d-1}, TMAX_{d-1},TMIN_d,TMAX_d,dt_{1d},...,dt_{6d}) \tag{3.3.57}$$

gdzie:

ZD_d – dobowe zapotrzebowanie na energię w dniu d,
$ZG_{d-1}(1)$, ..., $ZG_{d-1}(24)$ – godzinowy rozkład zużycia w dniu poprzednim,
$TMIN_{d-1}$, $TMAX_{d-1}$ – temperatura minimalna i maksymalna w dniu poprzednim,
$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,
dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Przypomnijmy, że model (3.3.57) dyskutowany był już w poprzednim rozdziale, w punkcie 2.3.2 (model 1). Przedziały prognozy przetestowano w tym przypadku dla predyktorów otrzymanych przy użyciu warstwowej sieci perceptronowej MLP oraz sieci neuronowo-rozmytej FBF (Bartkiewicz 2011a; Bartkiewicz 2012). Podobnie jak w poprzednim przypadku, porównaliśmy

częstości empiryczne przedziałów otrzymanych za pomocą bootstrapu i metody delta (patrz punkt 3.3.2, tabela 3.3.2). Wyniki porównania obu metod znajdują się w tabeli 3.3.7. Jak widzimy, również i w tym przypadku procent obserwacji rzeczywistego dobowego zapotrzebowania na energię w obrębie przedziałów prognozy otrzymanych metodą delta i przy użyciu bootstrapu kształtują się na zbliżonym poziomie. Częstości empiryczne oszacowań przedziałów dla obydwu podejść, zarówno w przypadku sieci neuronowej MLP, jak i neuronowo-rozmytej FBF, w wysokim stopniu odpowiadają oczekiwanym prawdopodobieństwom, dla których je wyznaczono. Również i tym razem można więc wysnuć wniosek o poprawnym efekcie wykorzystania podejścia empirycznego opartego na bootstrapie do oszacowania wariancji prognozy zapotrzebowania na energię elektryczną, zwracając jednak uwagę na fakt, że osiągane wyniki mają charakter zbliżony do tych otrzymanych w przypadku podejścia analitycznego, w którym wykorzystano metodę delta.

Tabela 3.3.7. Częstości empiryczne przedziałów prognozy dobowego zapotrzebowania na energię, otrzymanych za pomocą metody delta i bootstrapu, dla sieci MLP i FBF (w %)

Prawdopodobieństwo	Metoda delta		Bootstrap	
	MLP	FBF	MLP	FBF
80	81,24	81,77	81,84	80,06
85	85,63	86,16	85,80	84,36
90	89,57	90,25	89,86	89,34
95	93,99	93,25	93,99	93,49

Źródło: opracowanie własne.

Następnym zagadnieniem prognostycznym z zakresu krótkoterminowych prognoz obciążenia sieci elektroenergetycznej, dla którego badaliśmy dokładność przedziałów prognozy otrzymanych za pomocą bootstrapu, będzie prognoza zapotrzebowania na moc w szczycie wieczornym, z półdniowym wyprzedzeniem czasowym (model 3 przedstawiony w punkcie 2.3.2):

$$ZSW_d = f(ZSR_{d-1}, ZSW_{d-1}, TR_{d-1}, TP_{d-1}, TW_{d-1}, ZSR_d, TR_d, TP_d, TW_d) \tag{3.3.58}$$

gdzie:

ZSW_d – zapotrzebowanie na energię w szczycie wieczornym, w dniu d,
ZSR_d – zapotrzebowanie na energię w szczycie porannym, w dniu d,
TR_d – temperatura poranna (mierzona o godzinie 8), w dniu d,
TP_d – temperatura w południe (mierzona o godzinie 13), dniu d,
TW_d – temperatura wieczorna (mierzona o godzinie 21), w dniu d.

W tym przypadku jako model prognostyczny wykorzystaliśmy sieć neuronowo-rozmytą FBF. Podobnie jak w poprzednio rozważanych zagadnieniach, porównaliśmy działanie metody szacowania wariancji prognozy szczytowego zapotrzebowania opartej na bootstrapie z wynikami otrzymywanymi metodą delta. Częstości otrzymane dla przedziałów prognozy wyznaczonych dla sieci FBF za pomocą obu podejść przedstawione zostały w tabeli 3.3.8.

Tabela 3.3.8. Częstości empiryczne przedziałów prognozy mocy w szczycie wieczornym, otrzymanych dla sieci FBF za pomocą metody delta i bootstrapu (w %)

Prawdopodobieństwo	Metoda delta	Bootstrap
	FBF	FBF
80	81,23	81,06
85	85,83	85,68
90	89,80	89,64
95	93,64	93,94

Źródło: opracowanie własne.

Jak widzimy, również i w tym przypadku prezentowane w tabeli 3.3.8 częstości empiryczne otrzymanych przedziałów prognozy potwierdzają nasze poprzednio poczynione obserwacje na temat przydatności bootstrapu jako metody szacowania wariancji prognozy zapotrzebowania na energię. Testowanie w przypadku neuronowo-rozmytego predyktora w postaci sieci FBF, podobnie jak w innych rozważanych sytuacjach, daje zbliżone wyniki do oczekiwanych wartości prawdopodobieństw, porównywalne również do przedziałów prognozy szczytowego zapotrzebowania na energię, otrzymanych metodą delta w punkcie 3.3.2 (tabela 3.3.3) (Bartkiewicz 2011b; Bartkiewicz 2012).

Ostatnim zagadnieniem, dla którego badaliśmy wykorzystanie bootstrapu jako metody szacowania wariancji prognozy zapotrzebowania na energię, dla danego wzorca wejściowego $\mathbf{x}$ i przedziałów prognozy otrzymanych na jej podstawie, był problem predyktora w formie sieci neuronowo-rozmytej typu Takagi–Sugeno, z liniowymi następnikami reguł (Bartkiewicz 2011b; Bartkiewicz 2012). Przypomnijmy, że tego typu modele wykorzystywaliśmy w poprzednim rozdziale (punkt 2.3.4) do prognozy szczytowego zapotrzebowania na energię (maksymalnego godzinnego zapotrzebowania na energię) z dwudniowym wyprzedzeniem czasowym:

$$ZS_d = f(ZS_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZS(t)_{d-2}, ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \tag{3.3.59}$$

gdzie oznaczenia są analogiczne jak w poprzednio prezentowanych zagadnieniach:

$ZS(t)_d$ – zapotrzebowanie na energię w szczytowej godzinie, w dniu d,

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d, gdzie t oznacza godzinę szczytową,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Podobnie jak poprzedno rozważane zagadnienia, problem wyznaczania przedziałów prognozy dla modelu (3.3.59) rozważaliśmy również w punkcie 3.3.2. Przypomnijmy jednak, że analizowana w tym podrozdziale metoda delta oparta jest na analizie błędu kwadratowego. Może więc być stosowana w przypadku sieci neuronowo-rozmytej typu Takagi–Sugeno, pod warunkiem jednak, że wszystkie wagi modelu dopasowywane są metodą wstecznej propagacji minimalizującej błąd kwadratowy na zbiorze treningowym. Tymczasem w przypadku naszego predyktora nie korzystaliśmy z metody najmniejszych kwadratów, tylko z dwuetapowej procedury treningowej polegającej na oszacowaniu parametrów zbiorów rozmytych poprzedników za pomocą algorytmu grupowania danych, a następnie na wyznaczeniu współczynników funkcji liniowych następników reguł systemu metodą regresji liniowej (patrz punkt 2.3.4).

Tabela 3.3.9. Częstości empiryczne przedziałów prognozy maksymalnej energii godzinnej, otrzymanych za pomocą metody delta i bootstrapu dla sieci typu Takagi–Sugeno (w %)

Prawdopodobieństwo	Metoda delta	Bootstrap
80	76,95	77,11
85	83,53	83,69
90	88,25	88,11
95	92,49	92,49

Źródło: opracowanie własne.

Aby wyznaczyć wariancję prognozy przy danym wzorcu wejściowym **x** dla wykorzystywanej przez nas sieci neuronowo-rozmytej Takagi–Sugeno w punkcie 3.3.2 potraktowaliśmy ją jako uogólniony model regresji krzywoliniowej, uwzględniając wyłącznie niepewność współczynników funkcji liniowych, następników reguł. Pomimo tego otrzymane wyniki nie były wcale złe (tabela

3.3.4), co sugerowałoby, że poczynione uproszczenia nie miały takiego istotnego wpływu na oszacowanie wariancji prognozy zapotrzebowania na energię. Tym niemniej interesujące będzie porównanie z wartościami otrzymanymi za pomocą bootstrapu.

Oszacowanie wariancji prognozy z wykorzystaniem bootstrapu opiera się wyłącznie na obserwacji zachowania modelu i nie wymaga żadnych założeń odnośnie do jego architektury czy sposobu uczenia. Jak widzimy, analizując porównanie częstości dla przedziałów prognozy otrzymanych za pomocą obu metod (tabela 3.3.9), zastosowanie bootstrapu nie wniosło jakiejś specjalnej zmiany. Wyniki w obydwu przypadkach są bardzo zbliżone.

W bieżącym punkcie przedstawiliśmy wykorzystanie do oszacowania wariancji prognozy zapotrzebowania na energię elektryczną, dla danego wzorca wejściowego $\mathbf{x}$, podejścia opartego na bootstrapowaniu zbioru danych treningowych poprzez próbkowanie par. Metoda ta została przebadana dla kilku modeli neuronowych i neuronowo-rozmytych oraz szeregu zadań z zakresu krótkoterminowego prognozowania obciążenia sieci. Wyniki badań przedziałów prognozy wyznaczonych na podstawie oszacowanego rozkładu zapotrzebowania wskazują, że bootstrap może być jednym z narzędzi wykorzystywanych do rozwiązywania rozważanego zadania. Otrzymywane przedziały prognozy wskazują na uzyskanie dobrego przybliżenia szukanego rozkładu obciążenia.

Wykorzystanie podejścia opartego na bootstrapie w zagadnieniach krótkoterminowej prognozy zapotrzebowania na energię w świetle prezentowanych badań daje jednak w przypadku rozważanych modeli neuronowych i neuronowo-rozmytych zbliżone wyniki jak zastosowanie omawianej w punkcie 3.3.2 metody delta z pełnym oszacowaniem hesjanu błędu. Należy przy tym zwrócić uwagę, że odbywa się to kosztem naprawdę poważnych nakładów obliczeniowych, niezbędnych do uzyskania oszacowania wariancji prognozy za pomocą bootstrapu.

W naszych eksperymentach, jak już wspomnieliśmy, w każdym przedstawionym przypadku liczba bootstrapowanych prób we wzorze (3.3.51) wynosiła 30 ($B = 30$). Oznacza to konieczność budowy (nauczenia) za każdym razem 30 sieci neuronowych lub neuronowo-rozmytych. Zwłaszcza w przypadku modeli trenowanych metodą wstecznej propagacji błędu, która ma charakter iteracyjny i sama w sobie wymaga znacznych nakładów obliczeniowych, może to stanowić barierę zastosowań metod opartych na bootstrapie, tym bardziej że dla tego typu modeli możemy do wyznaczania wariancji prognozy z powodzeniem (jak wskazują badania w punkcie 3.3.2) stosować dużo mniej kosztowną metodę delta.

W przypadku sieci typu Takagi–Sugeno uczonych za pomocą dwuetapowej procedury treningowej, która polega na zastosowaniu algorytmu grupowania danych, a następnie regresji liniowej (patrz punkt 2.3.4), zastosowanie metody delta staje się bardziej problematyczne (aczkolwiek w testowanym przypadku

zakończyło się sukcesem). Ponieważ ta forma treningu nie wymaga tylu obliczeń, tworzenie grupy modeli na bootstrapowanych próbach jest szybsze i może być łatwiej zaakceptowane jako praktyczne rozwiązanie.

Pamiętać również należy, że ostateczny wybór metody szacowania wariancji prognozy powinien być dokonany dla konkretnego danego zadania prognostycznego. Choć być może z powodu dużych kosztów tworzenia podejście oparte na bootstrapie powinno być traktowane jako „drugi wybór", to jednak w świetle przedstawionych badań widzimy, że oferuje ono realną alternatywę w stosunku do metody delta. Nie można zapomnieć, że podejście to jest dużo bardziej uniwersalne i odporne na naruszenia założeń czynionych w przypadku metody analitycznej. Pomijając nawet trywialne sytuacje, że dla danego predyktora po prostu nie potrafimy wyznaczyć macierzy kowariancji parametrów albo (jak w przypadku rozważanych w tej pracy sieci typu Takagi–Sugeno) uzyskane oszacowanie może budzić wątpliwości, należy podkreślić, że zastosowanie bootstrapu daje nadzieję na sukces nawet gdy niepowodzeniem zakończyło się użycie metody delta. Być może linearyzacja modelu obarczona jest zbyt dużym błędem albo nie jest spełnione któreś z innych założeń, związanych z poprawnością modelu, wariancją błędu losowego itp. Oszacowanie wariancji prognozy zapotrzebowania oparte na bootstrapie, które nie zależy od tego typu założeń, może dać poprawne wyniki.

3.4. Modelowanie wariancji prognozy wynikającej z błędu losowego

3.4.1. Błąd losowy i błąd prognozy

Obecnie zajmiemy się zagadnieniem modelowania błędu losowego modelu prognostycznego. Interesować nas będą przy tym dwie podstawowe kwestie: znalezienie sposobu wyznaczania wariancji (lub odchylenia standardowego) elementu losowego oraz zweryfikowanie jego rozkładu prawdopodobieństwa. W obydwu przypadkach niezbędne informacje otrzymuje się na podstawie analizy reszt modelu otrzymanych dla jego zbioru treningowego.

Przypomnijmy bowiem, że zgodnie z rozważaniami przedstawionymi w punkcie 3.2.1 (patrz wzory (3.2.1) i (3.2.2)), wariancja prognozy zapotrzebowania na energię elektryczną (wyjścia modelu) $\sigma^2_y(\mathbf{x})$, dla danego wzorca wejściowego prognozy $\mathbf{x}$, określana jest przez dwa podstawowe elementy: wariancję wyjścia modelu prognostycznego wynikającą z niepewności parametrów (wag) $\sigma^2_{\mathrm{w}}(\mathbf{x})$ oraz właśnie wariancję błędu losowego. Jeżeli występuje również niepewność danych wejściowych prognozy, to należy uwzględnić ten

czynnik w postaci dodatkowego komponentu wariancji $\sigma^2_x(\mathbf{x})$. Określenie skali zmienności błędu losowego jest więc elementem niezbędnym do oszacowania rozkładu prognozy zapotrzebowania.

Drugą istotną kwestię stanowi określenie rozkładu prawdopodobieństwa błędu losowego. W zasadzie wszystkie nasze poprzednie rozważania w tym rozdziale były prowadzone przy założeniu, że niepewność prognozowanego zapotrzebowania na energię, dla danego wzorca wejściowego $\mathbf{x}$, opisana jest rozkładem normalnym $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$ o wartości oczekiwanej określonej przez wyjście modelu (prognozę) $f(\mathbf{x}, \mathbf{w})$ i odchyleniu standardowym $\sigma_y(\mathbf{x})$. Jak dyskutowaliśmy w punkcie 3.2.4 poświęconym niepewności wyjścia modeli liniowych, warunkiem przyjęcia założenia o normalności rozkładu wyjścia predyktora, jest normalny charakter rozkładu prawdopodobieństwa błędu losowego. W przypadku modeli nieliniowych, takich jak sieci neuronowe czy neuronowo-rozmyte, nawet w takiej sytuacji nie mamy gwarancji otrzymania właściwego kształtu rozkładu prognozy, ale w przybliżeniu możemy się tego spodziewać. A już z pewnością możemy powiedzieć, że normalny charakter rozkładu błędu losowego stanowi warunek konieczny dla normalności rozkładu prawdopodobieństwa prognozy zapotrzebowania, dla danego wzorca wejściowego $\mathbf{x}$.

Tak jak już wcześniej wspominaliśmy, modelowanie błędu losowego odbywa się na bazie analizy procesu resztowego, na zbiorze treningowym. Ponieważ reszty mierzone są dla tych samych wzorców danych, które wykorzystane zostały do dopasowania modelu, nie zawierają więc w sobie błędu generalizacji. Wobec tego, zakładając, że sam model jest poprawny, czyli, mówiąc w skrócie – jego wyjście stanowi dobrą aproksymację wartości oczekiwanej prognozy, rozkład reszt modelu, jako taki, stanowić będzie oszacowanie rozkładu prawdopodobieństwa błędu losowego zależności między zmiennymi wejściowymi a wyjściową, nieuwzględniające niepewności samego modelu.

Zanim przejdziemy dalej, poczyńmy jeszcze pewne zastrzeżenie. W naszej pracy rozważamy jedynie zagadnienie szacowania rozkładu reszt jako niezbędny element prezentowanej przez nas tematyki. Nie będziemy przedstawiać pełnej analizy procesu resztowego. W przypadku sieci neuronowych czy neuronowo-rozmytych formalna diagnostyka reszt nie jest tak istotna i bywa dużo rzadziej wykonywana niż dla modeli liniowych. Ponadto zasadniczo kwestie techniki tworzenia modelu prognostycznego, jego specyfikacji, uczenia i weryfikacji, pozostają poza obszarem naszych zainteresowań, jako wielokrotnie prezentowane w literaturze podstawowej, przede wszystkim z dziedziny sieci neuronowych. Również w zakresie diagnostyki reszt modelu zainteresowanych Czytelników odesłać możemy np. do publikacji Zapranis, Refenes 1999. Metody stosowane dla sieci neuronowo-rozmytych będą miały bardzo zbliżony charakter.

3.4.2. Czynnik losowy o stałym odchyleniu standardowym

Analizując rozkład warunkowy prognozy dla danego wzorca wejściowego, w przypadku modeli liniowych (punkt 3.2.4) zakładaliśmy, że reszty modelu (jego błąd losowy) mają rozkład $N(0, \sigma_\varepsilon)$, o wartości oczekiwanej 0 i stałym (niezależnym od wejścia) odchyleniu standardowym σ_ε. Przy podobnym założeniu wyprowadzana była metoda delta szacowania wariancji prognozy spowodowanej niepewnością wag $\sigma^2_{\mathbf{w}}(\mathbf{x})$ – w punkcie 3.3.2. Odchylenie standardowe σ_ε wyznaczane jest wówczas podobnie jak dla modeli liniowych, tj. przy użyciu standardowego estymatora błędu standardowego w próbie, czyli pierwiastka średniego kwadratu reszt modelu S_{N-p}.

Dla zbioru danych treningowych $D = \{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$ oszacowanie odchylenia standardowego σ_ε dane jest przez:

$$\sigma_\varepsilon = S_{N-p} = \sqrt{\frac{1}{N-p}\sum_{k=1}^{N} e_k^2} = \sqrt{\frac{1}{N-p}\sum_{k=1}^{N}(y_k - f(\mathbf{x}_k, \mathbf{w}))^2} \tag{3.4.1}$$

gdzie N jest licznością zbioru treningowego, p liczbą parametrów modelu (wag sieci neuronowej lub neuronowo-rozmytej).

W literaturze poświęconej sieciom neuronowym wskazuje się, że twórcy modeli neuronowych mają tendencję do pewnego przeszacowywania rozmiarów wykorzystywanych sieci i, w konsekwencji, stosowania zbyt dużej, redundantnej liczby wag. W związku z tym postuluje się czasami zastąpienie w (3.4.1) liczby parametrów p liczbą parametrów dobrze określonych, efektywnych p_{eff}:

$$\sigma_\varepsilon = \sqrt{\frac{1}{N-p_{eff}}\sum_{k=1}^{N} e_k^2} \tag{3.4.2}$$

Metoda szacowania liczby parametrów efektywnych p_{eff} została wprowadzona przez Moody'ego i określona jest ona przez ślad tzw. macierzy informacyjnej **G** (Moody 1992; Zapranis, Refenes 1999):

$$p_{eff} = \operatorname{tr} \mathbf{G} = \operatorname{tr} \mathbf{T}\mathbf{H}^{-1}\mathbf{T}^T \tag{3.4.3}$$

gdzie tr **G** oznacza ślad macierzy **G** (sumę elementów na głównej przekątnej), **H** jest macierzą hesjanu błędu treningowego E (patrz punkt 3.3.2), zaś macierz **T** składa się z pochodnych błędu E względem wyjścia sieci dla każdego wzorca treningowego $o_k = f(\mathbf{x}_k, \mathbf{w})$ oraz poszczególnych wag w_j:

$$T = [t_{kj}] = \left[\frac{\partial E}{\partial o_k \partial w_j}\right], \quad k = 1,\ldots N, j = 1,\ldots, p \tag{3.4.4}$$

Jak widzimy, oszacowanie (3.4.3) liczby parametrów efektywnych p_{eff}, pomimo skromnego zapisu, nie jest wcale proste i wymaga wielu obliczeń. Ponadto w badanych przez nas zagadnieniach liczba otrzymywana wag efektywnych była równa (bądź bardzo zbliżona) do liczby wszystkich wag p, a różnica między oszacowaniami odchylenia standardowego σ_ε w przypadku p_{eff} i p – nieznacząca. Wydaje się więc, że przy rozsądnym podejściu do budowy modelu oraz dużych, kilkusetelementowych próbach, z jakimi mamy do czynienia w zagadnieniach krótkoterminowego prognozowania zapotrzebowania na energię elektryczną, możemy poprzestać na oszacowaniu σ_ε w postaci (3.4.1).

Niestety w badanych przez nas przypadkach analiza reszt modelu nie potwierdzała założenia o rozkładzie normalnym błędu losowego $N(0, \sigma_\varepsilon)$, o stałym odchyleniu standardowym σ_ε. Przyjrzymy się dokładniej temu zjawisku na przykładzie zastosowania sieci neuronowej do prognozy maksymalnego godzinnego zapotrzebowania na energię, z dwudniowym wyprzedzeniem czasowym. Model zdefiniowany jest przez równanie postaci:

$$\begin{aligned} ZS_d = f(&ZS_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZS(t)_{d-2}, \\ &ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \end{aligned} \tag{3.4.5}$$

gdzie:

$ZS(t)_d$ – zapotrzebowanie na energię w szczytowej godzinie, w dniu d,

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d, gdzie t oznacza godzinę szczytową,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Wykorzystany zbiór treningowy obejmował roczny zbiór obserwacji procesu godzinnego zapotrzebowania energii oraz pomiarów temperatur, w jednej ze spółek dystrybucyjnych. Wykorzystywane dane obejmowały wszystkie dni tygodnia (włączając soboty i niedziele), ale usunięto z nich obserwacje dotyczące dni nietypowych, tj. świąt narodowych i religijnych przypadających w normalne robocze dni tygodnia, oraz dni poświątecznych (tak jak to dyskutowano w punkcie 2.2.5). Zbiór treningowy liczył więc ostatecznie 329 wzorców danych ($N = 329$).

Naszym celem jest zweryfikowanie założenia, że reszty treningowe modelu:

$$e_k = y_k - f(\mathbf{x}_k, \mathbf{w}), \quad k = 1, ..., N \tag{3.4.6}$$

mają rozkład normalny $N(0, \sigma_\varepsilon)$, o wartości oczekiwanej 0 i odchyleniu standardowym σ_ε, stałym i jednolitym w całej przestrzeni wejść (Zieliński 2000).

Tabela 3.4.1. Podstawowe statystyki reszt treningowych modelu e_k

Średnia	262,2	Kurtoza	1,149
Mediana	–120	Asymetria	0,286
Wariancja	1,98E+8	Minimum	–33 240
Odchylenie standardowe	14 072	Maksimum	45 720

Źródło: opracowanie własne.

Pierwszym krokiem, jaki wykonaliśmy w celu oceny rozkładu reszt (3.4.6), było wyznaczenie dla nich podstawowych statystyk z próby. Oszacowania ich wartości zostały przedstawione w tabeli 3.4.1.

Analizę statystyk z tabeli 3.4.1 rozpoczniemy od średniej z próby. Jak widzimy, średnia dla wszystkich 329 reszt wynosi 262,2, jest więc większa od zera. Powstaje oczywiście pytanie, czy oznacza to również, że wartość oczekiwana rozkładu prawdopodobieństwa reszt statystycznie istotnie różni się od zera. W celu weryfikacji tego faktu stosujemy standardowy test wartości oczekiwanej średniej z próby, przyjmując hipotezę zerową o zerowej wartości średniej. Statystyka testowa t wynosi w tym przypadku:

$$t = \frac{\bar{e}}{S_{\bar{e}}} = \frac{\bar{e}}{S_e}\sqrt{N} \approx 0{,}338 \tag{3.4.7}$$

Wartość krytyczna rozkładu t-Studenta dla tej statystyki, przy poziomie istotności $\alpha = 0{,}05$, wynosi w przybliżeniu 1,967, co nie pozwala nam na odrzucenie hipotezy zerowej o zerowej średniej. Ponadto weryfikacja tej decyzji okazuje się dosyć jednoznaczna, ponieważ poziom prawdopodobieństwa (istotności), który pozwalałby nam odrzucić hipotezę o zerowej wartości oczekiwanej, jest dosyć wysoki i wynosi $\alpha = 0{,}735$.

Z powodu pewnej asymetrii widocznej w rozkładzie reszt (współczynnik asymetrii wynoszący 0,286 i mediana nieco mniejsza niż średnia) zdecydowaliśmy się wykonać parametryczny test Lina i Mudholkara normalności rozkładu, przy asymetrycznej hipotezie alternatywnej. Statystykę L–M wykorzystywaną w tym teście możemy wyznaczyć w następujący sposób (Tong 1990):

– dla każdej reszty e_k wyznaczamy wartość roboczej statystyki z_k:

$$z_k = \left\{ \frac{1}{N}\left(\sum_{j\neq k}^{N} e_j^2 - \frac{1}{N-1}\left(\sum_{j\neq k}^{N} e_j\right)^2 \right) \right\}^{\frac{1}{3}}, \quad k = 1,\ldots,N \tag{3.4.8}$$

– wyznaczamy wartość statystyki L–M jako:

$$\mathrm{L-M} = \frac{1}{2}\sqrt{\frac{N}{3}}\ln\left(\frac{1+r_{ez}}{1-r_{ez}}\right) \tag{3.4.9}$$

gdzie r_{ez} jest współczynnikiem korelacji liniowej (Pearsona) pomiędzy serią reszt e_k i statystyk z_k, $k = 1, \ldots, N$.

Jeżeli hipoteza zerowa o normalności rozkładu reszt jest prawdziwa, to rozkładem asymptotycznym statystyki L–M jest rozkład normalny standardowy, o zerowej średniej i jednostkowym odchyleniu standardowym. Niestety test ten nie dał jednoznacznych wyników. Wartość statystyki L–M wynosiła w naszym przypadku –1,71. Wartość krytyczna rozkładu normalnego standardowego, pozwalająca na odrzucenie hipotezy zerowej odnośnie do normalności rozkładu i przyjęciu hipotezy alternatywnej o dodatniej asymetrii, na poziomie prawdopodobieństwa $\alpha = 0{,}05$, równa jest 1,64, czyli powinniśmy hipotezę normalności odrzucić. Poziom istotności pozwalający na zmianę tej decyzji jest jednak jedynie nieznacznie mniejszy, wynosi bowiem $\alpha = 0{,}0436$, tak więc test dał wyniki diagnostycznie niejednoznaczne.

Patrząc na niejednoznaczne wyniki testu Lina i Mudholkara oraz na widoczną kurtozę rozkładu (patrz tabela 3.4.1), zdecydowaliśmy się wykonać nieparametryczne testy zgodności rozkładu reszt z rozkładem normalnym. Istnieje cały szereg różnorodnych testów statystycznych, które można zastosować w celu weryfikacji hipotezy normalności rozkładu – stanowią one wyniki odmiennych podejść do tego problemu. Wymienić można tutaj choćby test Shapiro–Wilka – wynikający z analizy wykresu kwantyli badanego rozkładu względem kwantyli testowanego rozkładu normalnego, test Jarque–Bery – porównujący wybrane momenty badanego rozkładu z momentami testowanego rozkładu normalnego, testy Kołmogorowa–Smirnowa i chi-kwadrat – porównujące empiryczne rozkłady reszt (w formie częstości bądź dystrybuanty empirycznej) z wartościami oczekiwanymi dla testowanego rozkładu normalnego. W przypadku diagnostyki reszt, przy dużej liczbie badanych wzorców, wybór metody weryfikacji w zasadzie jest nieco arbitralny. Na potrzeby obecnej analizy przyjrzymy się bliżej dwóm ostatnim klasycznym testom zgodności, Kołmogorowa–Smirnowa i chi-kwadrat.

Wykorzystanie obydwu testów w diagnostyce procesu resztowego, jak już nadmieniliśmy, polega na porównaniu empirycznego rozkładu otrzymanych reszt modelu prognostycznego z oczekiwanym rozkładem normalnym $N(0, \sigma_\varepsilon)$. Zazwyczaj wcześniej normalizuje się badane reszty:

$$u_k = \frac{e_k}{\sigma_\varepsilon},\ k = 1,\ldots,N \tag{3.4.10}$$

porównując je dalej z rozkładem normalnym standardowym $N(0, 1)$. Jest to procedura niezbędna w przypadku znajdowania wartości krytycznych rozkładów testowych za pomocą tablic statystycznych, ale można zastosować ją zawsze, nawet jeśli dysponujemy odpowiednim oprogramowaniem pozwalającym na ich obliczenie dla dowolnych wartości parametrów.

Pokrótce przedstawimy najważniejsze elementy obydwu procedur testowych.

Test Kołmogorowa–Smirnowa

Przy użyciu tego testu porównuje się dystrybuantę empiryczną rozkładu reszt z wartościami oczekiwanymi dla rozkładu $N(0, 1)$. Przypomnijmy, że o dystrybuancie empirycznej błędu wspominaliśmy już w punkcie 3.2.3. W naszym przypadku znormalizowane reszty modelu u_k porządkujemy w sposób niemalejący, otrzymując zbiór statystyk porządkowych rozkładu błędu $u_{(1)} \le u_{(2)} \le \ldots \le u_{(N)}$. Dystrybuantę empiryczną rozkładu błędu dla naszej N-elementowej próby reszt możemy zapisać następująco:

$$P_N(u) = \begin{cases} 0 & , u < u_{(1)} \\ \dfrac{k}{N} & , u_{(k)} \le u < u_{(k+1)} \\ 1 & , u_{(N)} \le u \end{cases} \tag{3.4.11}$$

W niektórych przypadkach (np. jeżeli dalej do znalezienia dystrybuanty rozkładu normalnego standardowego $N(0, 1)$ używać będziemy tablic statystycznych) rozłożenie punktów $u_{(k)}$ musi być przyjęte z góry, wtedy dystrybuantę empiryczną tworzymy klasycznie ze skumulowanych liczności (zliczeń) w szeregu rozdzielczym (histogramie) rozkładu reszt na predefiniowane podprzedziały.

Statystyką testową jest maksymalna odległość między wartościami dystrybuanty rozkładu reszt i standardowego rozkładu normalnego:

$$D_N = \max_k \left| P_N(u_{(k)}) - P_{N(0,1)}(u_{(k)}) \right| = \max_k \left| \frac{k}{N} - P_{N(0,1)}(u_{(k)}) \right| , k = 1,\ldots,N \tag{3.4.12}$$

Czasami, aby uniknąć problemów z przybliżaniem ciągłego rozkładu prawdopodobieństwa za pomocą zbioru dyskretnych punktów (to jest z domknięciami przedziałów w dystrybuancie empirycznej (3.4.11)), stosuje się bardziej rozbudowaną wersję statystyki D_N:

$$D_N = \max_k \left(\max \left(\left| P_N(u_{(k)}) - P_{N(0,1)}(u_{(k)}) \right|, \left| P_N(u_{(k-1)}) - P_{N(0,1)}(u_{(k)}) \right| \right) \right), k = 1,...,N \quad (3.4.13)$$

W przypadku dużej liczby reszt gęsto próbkujących rozkład stosowanie (3.4.12) nie powinno jednak powodować specjalnych problemów.

Jeżeli spełniona jest hipoteza zerowa o normalności rozkładu reszt, to wówczas statystyka D_N ma rozkład Kołmogorowa. Hipotezę zerową możemy odrzucić na poziomie prawdopodobieństwa (istotności) α, jeżeli:

$$\sqrt{N} D_N > K_\alpha \quad (3.4.14)$$

gdzie K_α jest wartością krytyczną testu dla tego prawdopodobieństwa. Musimy tutaj jednak uważać, ponieważ jeżeli wartości krytyczne tablicowane są (bądź obliczane przez stosowane oprogramowanie) z wykorzystaniem dwóch parametrów $K_{\alpha,N}$, oznacza to, że są one już zdenormalizowane dla rozmiaru próby i obszar odrzucenia definiowany jest jako:

$$D_N > K_{\alpha,N} \ , \qquad K_{\alpha,N} = \frac{K_\alpha}{\sqrt{N}} \quad (3.4.15)$$

Test chi-kwadrat

Test chi-kwadrat jest jednym ze standardowych testów zgodności wchodzących w skład podstawowych kursów statystyki. W diagnostyce normalności reszt wykorzystuje się go w formie porównania zgodności rozkładu ich liczności na pewien zbiór podprzedziałów z prawdopodobieństwami tychże podprzedziałów dla rozkładu normalnego.

Jeżeli więc mamy zbiór znormalizowanych reszt (3.4.10) u_k, $k = 1, \ldots, N$, to dzielimy przedział wartości wszystkich reszt na L rozłącznych, sąsiadujących podprzedziałów, których krańce określone są przez uporządkowany zbiór punktów $u_{(1)} \le u_{(2)} \le \ldots \le u_{(L)}$, gdzie oczywiście $L < N$. Tworzymy szereg rozdzielczy, wyznaczając liczności (zliczenia) reszt w poszczególnych podprzedziałach:

$$n_l = \begin{cases} \left| \{u_k : u_k \le u_{(1)}\} \right| & , l = 1 \\ \left| \{u_k : u_{(l-1)} < u_k \le u_{(l)}\} \right| & , l = 2,...,L \end{cases} , \ k = 1,...,N \quad (3.4.16)$$

gdzie operacja $|\cdot|$ oznacza moc (liczbę elementów) zbioru. Oczywiście suma wszystkich liczności n_l, $l = 1, \ldots, L$ musi być równa liczbie wszystkich reszt N.

Dalej, korzystając z dystrybuanty rozkładu normalnego standardowego $P_{N(0,1)}$, wyznaczamy prawdopodobieństwa poszczególnych przedziałów:

$$p_l = \begin{cases} P_{N(0,1)}(u_{(1)}) & , l = 1 \\ P_{N(0,1)}(u_{(l)}) - P_{N(0,1)}(u_{(l-1)}) & , l = 2,...,L \end{cases} \tag{3.4.17}$$

a następnie – oczekiwane liczności teoretyczne o_l, $l = 1, \ldots, L$ reszt w poszczególnych podprzedziałach:

$$o_l = Np_l \quad , l = 1,...,L \tag{3.4.18}$$

Statystykę testową *ch* wyznaczamy w następujący sposób:

$$ch = \sum_{l=1}^{L} \frac{(o_l - n_l)^2}{o_l} \tag{3.4.19}$$

ponieważ oszacowanie wariancji dla zliczeń wynosi również o_l.

Hipotezę zerową o normalności badanego rozkładu reszt możemy odrzucić na poziomie prawdopodobieństwa (istotności) α, jeżeli spełniony jest warunek:

$$ch > \chi^2(\alpha, L-2) \tag{3.4.20}$$

gdzie χ^2 jest wartością krytyczną rozkładu chi-kwadrat, dla poziomu prawdopodobieństwa α oraz L–2 stopni swobody (ponieważ testowany rozkład teoretyczny ma w tym przypadku dwa szacowane parametry).

Obydwa testy zostały przeprowadzone w przypadku reszt modelu neuronowego dla rozważanego przez nas w bieżącym punkcie zagadnienia prognozy szczytowego zapotrzebowania na energię godzinną (zależność (3.4.5)). Informacje dotyczące wyników działania obu procedur diagnostycznych przedstawione zostały w tabeli 3.4.2 (Zieliński 2000). Zawiera ona wartość statystyki testowej, odpowiednią wartość krytyczną rozkładu testowego na poziomie istotności α = 0,05 i poziom prawdopodobieństwa dla ewentualnej zmiany decyzji o odrzuceniu hipotezy zerowej.

Tabela 3.4.2. Wyniki testów normalności standaryzowanych reszt treningowych modelu u_k dla oszacowanej wartości σ_ε

Rodzaj testu	Statystyka testowa	Wartość krytyczna (α = 0,05)	Poziom istotności
Test chi-kwadrat	46,418	12,591	1E-08
Test K–S	1,8836	1,36	0,0017

Źródło: opracowanie własne.

Jak widzimy, wyniki przeprowadzonych testów są dosyć jednoznaczne. Wartości statystyk testowych w obydwu przypadkach okazały się zdecydowanie wyższe od odpowiednich wartości krytycznych. Pozwala nam więc to odrzucić hipotezę zerową o normalnym rozkładzie reszt badanego modelu neuronowego prognozy zapotrzebowania na energię na przyjętym poziomie istotności $\alpha = 0{,}05$. Ponadto decyzja ta ma charakter bardzo jednoznaczny. Jak widzimy w ostatniej kolumnie, hipotezę normalności możemy odrzucić na bardzo niskich poziomach prawdopodobieństwa popełnienia błędu.

Na koniec spójrzmy jeszcze na tabelę 3.4.3. Zawiera ona, dla kilku wybranych poziomów ε, stablicowane względne wartości liczby znormalizowanych reszt u_k spełniających warunek:

$$|u_k| \geq \varepsilon \tag{3.4.21}$$

Tabela 3.4.3. Procent znormalizowanych reszt treningowych modelu u_k, dla stałej wartości σ_ε, przekraczających podane poziomy wielkości ε

ε	0,5	0,75	1	1,25	1,5	1,75	2
$\lvert u_k \rvert \geq \varepsilon$	44,68	27,96	16,72	10,33	5,78	2,43	1,22
$N(0,1)$	61,71	45,33	31,73	21,13	13,36	8,01	4,55

Źródło: opracowanie własne.

Pozwoli nam to ocenić różne poziomy rozrzutu wartości bezwzględnej reszt wokół wartości oczekiwanej 0. W tabeli 3.4.3 zostały one skonfrontowane z odpowiednimi liczbami oczekiwanymi dla rozkładu normalnego standardowego.

Analizując dane w tabeli 3.4.3, widzimy, że rozkład badanych reszt wyraźnie odbiega od założonego rozkładu normalnego o stałym odchyleniu standardowym σ_ε wyznaczonym za pomocą zależności (3.4.2) (lub (3.4.1)). Rozrzut reszt zdecydowanie przekracza oczekiwane poziomy, co wskazuje w badanym przypadku na problemy z oszacowaniem wielkości odchylenia standardowego błędu losowego w konfrontacji z założeniem normalności jego rozkładu.

Podsumowując wnioski z bieżącego punktu, możemy powiedzieć, że w zadaniach krótkoterminowego prognozowania zapotrzebowania na energię elektryczną, standardowe metody szacowania odchylenia standardowego (wariancji) rozkładu prawdopodobieństwa czynnika losowego dla nieliniowych modeli predykcji, takich jak sieci neuronowe czy neuronowo-rozmyte, nie zawsze muszą dawać poprawne wyniki. Wykorzystanie do wyznaczenia σ_ε klasycznego oszacowania błędu standardowego z próby, za pomocą zależności (3.4.1) czy też (3.4.2) (w badanym przypadku obydwa dały identyczne wyniki), może nie pozwolić na podtrzymanie założenia o normalnym rozkładzie elementu losowe-

go, a co za tym idzie, o rozkładzie przewidywanego zapotrzebowania dla danego wzorca wejściowego prognozy. Jak pokażemy w punkcie następnym, powód może stanowić przyjęcie założenia o stałym charakterze odchylenia σ_ε.

3.4.3. Czynnik losowy o zmiennym odchyleniu standardowym

W poprzednim punkcie analizowaliśmy model błędu losowego predyktora neuronowego (neuronowo-rozmytego) w zadaniach prognozowania zapotrzebowania na energię elektryczną, przy założeniu normalności jego rozkładu $N(0, \sigma_\varepsilon)$ i stałej (niezależnej od wejścia prognozy) wartości odchylenia standardowego σ_ε. Jak wskazuje przedstawiony przykład, założenie to nie zawsze może być podtrzymane. Dlatego jeszcze raz zwróćmy tutaj uwagę, że jakiekolwiek wnioskowanie o niepewności każdego wykorzystywanego modelu prognostycznego powinno zostać poprzedzone analizą jego rozkładu reszt w celu weryfikacji przyjętego modelu rozkładu prawdopodobieństwa błędu.

Nasze dotychczasowe doświadczenia wskazują wręcz, że w przypadku zastosowań sieci neuronowych i neuronowo-rozmytych do krótkoterminowych prognoz obciążenia sieci, założenie o rozkładzie błędu losowego $N(0, \sigma_\varepsilon)$ często nie jest spełnione. Problem leży tutaj w przyjmowanym stałym charakterze odchylenia standardowego (wariancji) czynnika losowego σ_ε szacowanego za pomocą (3.4.1) lub (3.4.2). Znacznie lepsze wyniki otrzymaliśmy, wykorzystując odmienny model błędu losowego, zależnego od wzorca wejściowego prognozy $\mathbf{x}$.

W bieżącym punkcie będziemy więc zakładać, że błędy losowe modelu są niezależne od siebie, mają rozkład normalny o wartości oczekiwanej 0, ale jego odchylenie standardowe $\sigma_\varepsilon(\mathbf{x})$ nie jest stałe, lecz może zmieniać się w różnych obszarach przestrzeni wejść. Heteroskedastyczność (ponieważ tak nazywamy tę właściwość) nieznanej postaci nie jest zresztą niczym niezwykłym w przypadku modeli nieliniowych. Najważniejszy problem, jaki musimy tutaj rozwiązać, stanowi dostarczenie odpowiedniego oszacowania wartości $\sigma_\varepsilon(\mathbf{x})$. Model odchylenia standardowego $\sigma_\varepsilon(\mathbf{x})$ (lub wariancji $\sigma^2_\varepsilon(\mathbf{x})$) błędu losowego możemy otrzymać np. za pomocą dodatkowej sieci neuronowej uczonej z wykorzystaniem reszt oryginalnego modelu prognostycznego.

Jeżeli, tak jak w poprzednim punkcie, oznaczymy przez $D = \{\mathbf{x}_k, y_k\}$, $k = 1, \ldots, N$ zbiór danych treningowych wykorzystanych do budowy neuronowego lub neuronowo-rozmytego modelu prognozy zapotrzebowania na energię, to dodatkowa sieć neuronowa dostarczająca wartości odchylenia standardowego elementu losowego $\sigma_\varepsilon(\mathbf{x})$ może zostać stworzona przy użyciu następującego zbioru treningowego:

$$D_\varepsilon = \{\mathbf{x}_k, |e_k|\}, k = 1, \ldots, N \tag{3.4.22}$$

Jak widzimy, zbiór treningowy D_ε zbudowany jest z wzorców, które w części wejściowej składają się z wektora wejściowego $\mathbf{x}_k$, natomiast wartościami docelowymi w procesie treningu dodatkowej sieci, $|e_k| = |y_k - f(\mathbf{x}_k, \mathbf{w})|$, $k = 1, \ldots, N$, są wartości bezwzględne reszt treningowych modelu prognostycznego.

Wymiennie moglibyśmy również modelować wariancję błędu losowego $\sigma^2_\varepsilon(\mathbf{x})$, trenując dodatkową sieć neuronową na zbiorze danych:

$$D_\varepsilon = \{\mathbf{x}_k, (e_k)^2\},\ k = 1, \ldots, N \tag{3.4.23}$$

Wróćmy teraz do przykładu z poprzedniego punktu 3.4.2 dotyczącego zastosowania sieci neuronowej MLP do prognozy maksymalnego godzinnego zapotrzebowania na energię, z dwudniowym wyprzedzeniem czasowym (zależność (3.4.5)). Po negatywnych próbach modelowania błędu losowego przy stałym σ_ε sporządziliśmy model odchylenia standardowego reszt $\sigma_\varepsilon(\mathbf{x})$, w którym wykorzystuje się dodatkową warstwową sieć perceptronową, trenowaną zgodnie ze wzorem (3.4.22) na podstawie zależności:

$$\begin{aligned} |ZS_d - MLP| = f(&ZS_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZS(t)_{d-2}, \\ &ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, \ldots, dt_{6d}) \end{aligned} \tag{3.4.24}$$

gdzie MLP jest prognozą zapotrzebowania na energię z podstawowego modelu, inne oznaczenia tak samo jak w (3.4.5).

Aby zweryfikować to podejście i otrzymane oszacowanie odchylenia standardowego błędu losowego $\sigma_\varepsilon(\mathbf{x})$, przeprowadziliśmy diagnostykę normalności reszt wyjściowego modelu prognostycznego (3.4.5), przy założeniu, że oczekiwany rozkład empiryczny ma charakter rozkładu normalnego $N(0, \sigma_\varepsilon(\mathbf{x}))$. Podobnie jak w poprzednim punkcie, wykonaliśmy test Kołmogorowa–Smirnowa i chi-kwadrat. W tym przypadku, normalizując reszty, dzieliliśmy je przez odchylenie standardowe obliczone przez dodatkową sieć neuronową dla wejścia danego wzorca treningowego:

$$u_k = \frac{e_k}{\sigma(\mathbf{x}_k)}, \quad k = 1, \ldots, N \tag{3.4.25}$$

Jeżeli oszacowanie odchylenia standardowego reszt $\sigma_\varepsilon(\mathbf{x})$ jest poprawne, a rozkład błędu losowego – normalny, to znormalizowane (standaryzowane) reszty u_k, $k = 1, \ldots, N$ powinny wszystkie mieć standardowy rozkład normalny $N(0, 1)$. Wyniki przeprowadzonych testów przedstawione zostały w tabeli 3.4.4.

Jak widzimy, w tym przypadku wyniki testów normalności dosyć jednoznacznie nie pozwalają nam na odrzucenie hipotezy zerowej o zgodności rozkładu znormalizowanych reszt ze standardowym rozkładem normalnym.

Jeżeli spojrzymy na wartości statystyk testowych, to możemy zauważyć, że są one dużo mniejsze od odpowiadających im wartości krytycznych rozkładów, na poziomie istotności $\alpha = 0{,}05$. Prawdopodobieństwa zmiany decyzji i odrzucenia hipotezy zerowej wymagałyby założenia dosyć wysokiego poziomu prawdopodobieństwa popełnienia błędu.

Podobnie jak w poprzednim punkcie 3.4.2, dla stałej wartości odchylenia standardowego σ_ε (patrz tabela 3.4.3), ocenimy teraz jeszcze kształt całego rozkładu prawdopodobieństwa znormalizowanych reszt modelu u_k, wykorzystując uzyskane oszacowanie $\sigma_\varepsilon(\mathbf{x})$. Odpowiednie procentowe udziały reszt odchylających się od wartości oczekiwanej 0 powyżej podanych poziomów ε zostały przedstawione w tabeli 3.4.5. Jak widzimy, przy stworzonym modelu zmieniającego się odchylenia standardowego błędu losowego $\sigma_\varepsilon(\mathbf{x})$, zależnego od wzorca wejściowego prognozy $\mathbf{x}$, uzyskany kształt rozkładu reszt jest bardzo zbliżony do wartości oczekiwanych dla rozkładu normalnego standardowego. Pozwala to nam intuicyjnie potwierdzić pozytywne wyniki przeprowadzonych wcześniej testów normalności.

Tabela 3.4.4. Wyniki testów normalności standaryzowanych reszt treningowych modelu u_k dla oszacowanego odchylenia standardowego $\sigma_\varepsilon(\mathbf{x})$

Rodzaj testu	Statystyka testowa	Wartość krytyczna ($\alpha = 0{,}05$)	Poziom istotności
Test chi-kwadrat	6,595	12,592	0,35995
Test K–S	0,8311	1,36	0,4807

Źródło: opracowanie własne.

Tabela 3.4.5. Procent znormalizowanych reszt treningowych modelu u_k, odchylenia standardowego reszt $\sigma_\varepsilon(\mathbf{x})$, przekraczających podane poziomy wielkości ε

ε	0,5	0,75	1	1,25	1,5	1,75	2
$\lvert u_k \rvert \geq \varepsilon$	59,57	39,21	26,14	16,72	11,85	6,69	3,65
$N(0,1)$	61,71	45,33	31,73	21,13	13,36	8,01	4,55

Źródło: opracowanie własne.

Jak widzimy, zastąpienie stałej wartości odchylenia standardowego σ_ε zmieniającym się, zależnym od wzorca wejściowego $\mathbf{x}$ oszacowaniem $\sigma_\varepsilon(\mathbf{x})$, które otrzymuje się za pomocą dodatkowej sieci neuronowej (3.4.24), pozwoliło nam uzyskać korzystniejszy, lepiej odpowiadający założeniom, kształt rozkładu prawdopodobieństwa błędu losowego $N(0, \sigma_\varepsilon(\mathbf{x}))$. W wyraźny sposób przekłada się to również na ocenę rozkładu całej prognozy zapotrzebowania na energię.

Dla analizowanego w tym podrozdziale przykładu zastosowania sieci neuronowej MLP do prognozy maksymalnego godzinnego zapotrzebowania na

energię, z dwudniowym wyprzedzeniem czasowym (patrz wzór (3.4.5) w poprzednim podpunkcie), przetestowaliśmy oszacowanie odchylenia standardowego przewidywanej wartości popytu $\sigma_y(\mathbf{x})$, uzyskane dla stałego σ_ε (otrzymanego za pomocą wzoru (3.4.1)) oraz zmiennego σ_ε, określonego przez dodatkową sieć neuronową (3.4.24). W obydwu przypadkach zmienność modelu wynikająca z niepewności jego parametrów $\sigma_{\mathbf{w}}(\mathbf{x})$ wyznaczona została przy użyciu metody delta (omawianej w punkcie 3.3.2). Weźmiemy więc pod uwagę dwa oszacowania odchylenia standardowego finalnej prognozy:

– w pierwszym przypadku, dla stałego σ_ε:

$$\sigma_y(\mathbf{x}) = \sigma_\varepsilon \sqrt{1 + \mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{g}(\mathbf{x}, \mathbf{w}^*)} \tag{3.4.26}$$

– w przypadku drugim, dla oszacowania zależnego od $\mathbf{x}$, $\sigma_\varepsilon(\mathbf{x})$:

$$\sigma_y(\mathbf{x}) = \sigma_\varepsilon(\mathbf{x}) \sqrt{1 + \mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{g}(\mathbf{x}, \mathbf{w}^*)} \tag{3.4.27}$$

Przypomnijmy, że w (3.4.26) i (3.4.27) przez $\mathbf{H}$ oznaczyliśmy macierz hesjanu błędu treningowego modelu, zaś $\mathbf{g}(\mathbf{x}, \mathbf{w}^*)$ jest wektorem gradientu wyjścia sieci względem wag $\mathbf{w}$, dla wartości optymalnej (uzyskanej w wyniku uczenia) tych parametrów $\mathbf{w}^*$. Odnośnie do szczegółowych informacji odsyłamy Czytelnika do punktu 3.3.2.

Prognozę maksymalnego godzinnego zapotrzebowania na energię (3.4.5), uzyskaną z wykorzystaniem sieci neuronowej, przetestowano na ponaddwuletnim zbiorze danych $D_t = \{\mathbf{x}_t, y_t\}$, $t = 1, \ldots, T$. Liczba testowych wzorców danych, po odrzuceniu dni nietypowych (świąt i dni poświątecznych), wynosiła $T = 732$. Aby ocenić rozkład prognozy, wyznaczono zbiór znormalizowanych odchyleń testowych przewidywanego zapotrzebowania dla obu oszacowań odchylenia standardowego wyjścia modelu $\sigma_y(\mathbf{x})$, określonych wzorami (3.4.26) oraz (3.4.27):

$$u_t = \frac{y_t - f(\mathbf{x}_t, \mathbf{w})}{\sigma_y(\mathbf{x}_t)} \quad , \quad t = 1, \ldots, T \tag{3.4.28}$$

Podobnie jak w poprzednich przypadkach, w tabeli 3.4.6 przedstawiono stablicowane dla kilku wybranych poziomów ε względne liczności znormalizowanych odchyleń testowych prognozy zapotrzebowania u_t, wyznaczonych dla obu oszacowań odchylenia standardowego wyjścia modelu $\sigma_y(\mathbf{x})$, spełniających warunek:

$$|u_t| \geq \varepsilon \tag{3.4.29}$$

Analizując dane w tabeli 3.4.6, ponownie możemy dostrzec, że zastosowanie zmiennego oszacowania odchylenia standardowego błędu losowego $\sigma_\varepsilon(\mathbf{x})$, otrzymanego za pomocą dodatkowej sieci neuronowej (3.4.24), daje znacznie lepszy kształt rozkładu prawdopodobieństwa prognozy zapotrzebowania na energię, w dużo większym stopniu odpowiadający przyjętemu założeniu o jego normalności.

Dodajmy jeszcze, że wszystkie badania empiryczne przedziałów prognozy prezentowane w poprzednim podrozdziale 3.3 dla różnych oszacowań zmienności prognozy spowodowanej niepewnością parametrów wykonywane były z wykorzystaniem modelu odchylenia standardowego zależnego od wzorca wejściowego $\sigma_\varepsilon(\mathbf{x})$, który otrzymywano za pomocą dodatkowej sieci MLP trenowanej na zborze danych w formie (3.4.22). Prezentowane tam wyniki również potwierdzają dobre działanie tego podejścia w przypadku różnych badanych sieci neuronowych i neuronowo-rozmytych oraz wielu odmiennych zadań związanych z krótkoterminowym prognozowaniem zapotrzebowania na energię.

Tabela 3.4.6. Procent znormalizowanych reszt treningowych modelu u_k, odchylenia standardowego reszt $\sigma_\varepsilon(\mathbf{x})$, przekraczających podane poziomy wielkości ε

ε	0,5	0,75	1	1,25	1,5	1,75	2
$\lvert u_t \rvert \geq \varepsilon$, σ_ε	50,12	34,14	20,22	14,16	8,84	5,21	3,39
$\lvert u_t \rvert \geq \varepsilon$, $\sigma_\varepsilon(\mathbf{x})$	60,77	43,22	29,30	20,34	14,16	9,56	6,90
$N(0,1)$	61,71	45,33	31,73	21,13	13,36	8,01	4,55

Źródło: opracowanie własne.

3.5. Modelowanie niepewności wejść

3.5.1. Prognozowanie w warunkach szumu wejściowego

Jedną ze stałych przesłanek, które przyjmuje się dla modeli regresyjnych, stanowi założenie deterministycznego charakteru zmiennych objaśniających. Oznacza to, że również w przypadku rozważanych przez nas zagadnień krótkoterminowej prognozy zapotrzebowania na energię elektryczną dane wejściowe dla wykorzystywanego modelu neuronowego lub neuronowo-rozmytego traktowane są jako znane i pewne. Powstaje oczywiście pytanie, czy sytuacja taka odpowiada rzeczywistości.

Na tak zadane pytanie należy wyraźnie odpowiedzieć, że w ogólnym przypadku założenie o deterministycznym charakterze wejść modelu prognostyczne-

go jest często nierealistyczne. Dotyczy to może nie tyle procesu oszacowania jego parametrów (uczenia), co raczej predykcji na podstawie gotowego modelu. W tym pierwszym przypadku często bowiem dysponujemy dokładnymi historycznymi obserwacjami modelowanych procesów, które mogą być niedostępne w czasie normalnej eksploatacji modelu. Oto typowe przykłady takiej sytuacji:

– jako wejścia do modelu wykorzystujemy prognozy pewnych wielkości; dla danych treningowych albo prognoz testowych (*ex post*) możemy zwykle korzystać ze znanych wartości dokładnych; w przypadku praktycznej eksploatacji modelu, tworząc prognozy w nowych warunkach, musimy korzystać z niepewnych, obarczonych błędem przewidywań,

– w modelach szeregów czasowych często dokonujemy prognozy na kilka kroków do przodu; jako wejścia modelu wykorzystujemy wtedy zwykle prognozy z poprzednich kroków,

– pewne wejścia systemu, których wartości są znane w czasie treningu modelu, podczas sporządzania prognozy mogą być znane jedynie w przybliżeniu bądź też w ogóle niedostępne, muszą być więc wówczas aproksymowane.

Wszystkie opisane tu sytuacje są dosyć typowe dla wielu zadań prognostycznych. W każdym z rozważanych przypadków zmienne wejściowe dla prognozy nie mają charakteru deterministycznego, lecz są pewnymi zmiennymi losowymi, których wartości znane są nam jedynie z dokładnością do pewnego rozkładu prawdopodobieństwa. Problem niepewności wejść ma więc istotne znaczenie dla licznych zagadnień prognostycznych, z którymi mamy do czynienia w praktyce. Ewentualne występowanie procesu szumu zmiennych wejściowych traktowane musi być bowiem jako dodatkowy składnik losowy w realizacji zmiennej objaśnianej.

Z podobną sytuacją zetkniemy się również w odniesieniu do prognoz krótkoterminowego zapotrzebowania na energię elektryczną. Jeżeli przyjrzymy się przeglądowi zadań prognostycznych z tej dziedziny przedstawionemu w rozdziale 2, to stwierdzimy, że niemal w każdym przypadku wykorzystywane sieci neuronowe lub neuronowo-rozmyte wymagają podania jako danych wejściowych temperatur na prognozowany dzień. Podczas uczenia wykorzystywane są zarejestrowane dosyć dokładne wartości temperatur otrzymane ze zagregowanych danych z instrumentów pomiarowych. Na etapie operacyjnego wykorzystania modelu, w sposób naturalny, tego typu informacje dla okresu prognozy nie są dostępne. Musimy więc posługiwać się oszacowaniami temperatur na podstawie prognoz meteorologicznych.

Powstaje oczywiście pytanie, jak duży jest wpływ błędu predykcji temperatur na finalną prognozę zapotrzebowania na energię elektryczną. W tym celu przyjrzyjmy się porównaniu błędów uzyskanych dla dwóch, rozważanych już wcześniej w punkcie 2, zagadnień prognostycznych, które będziemy wykorzystywać w przykładach prezentowanych w bieżącym podrozdziale.

Pierwszym rozważanym zagadnieniem będzie, omawiane już wcześniej w punktach 2.2.4 i 2.2.5, zadanie prognozy godzinnego zapotrzebowania na energię elektryczną, z dwudniowym wyprzedzeniem czasowym:

$$ZG_d(t) = f(ZG(t)_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZG(t)_{d-2}, ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \quad (3.5.1)$$

gdzie oznaczenia są takie same jak w punkcie 2.2.4:

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Tabela 3.5.1. Porównanie dokładności prognozy godzinnego zapotrzebowania na energię dla dokładnych i zaszumionych wartości temperatur

Godz.	Temperatury dokładne		Prognozy temperatur		Godz.	Temperatury dokładne		Prognozy temperatur	
	MAE (kWh)	MAPE (%)	MAE (kWh)	MAPE (%)		MAE (kWh)	MAPE (%)	MAE (kWh)	MAPE (%)
1	6 262	2,99	6 164	2,95	13	6 429	2,57	7 824	3,12
2	4 863	2,39	5 050	2,50	14	5 600	2,25	7 528	3,00
3	4 438	2,24	4 645	2,35	15	6 450	2,52	8 253	3,20
4	3 970	2,01	4 373	2,22	16	9 196	3,60	10 467	4,10
5	4 130	2,06	4 500	2,25	17	8 385	3,19	9 478	3,61
6	4 544	2,21	4 890	2,38	18	7 918	2,89	8 523	3,05
7	6 292	2,71	6 576	2,85	19	7 482	2,69	7 625	2,71
8	6 642	2,71	7 003	2,86	20	4 909	1,61	5 341	1,75
9	6 239	2,51	6 639	2,67	21	5 822	1,99	6 407	2,18
10	6 631	2,61	7 092	2,79	22	6 203	2,36	6 356	2,42
11	7 153	2,81	7 873	3,10	23	5 078	2,07	5 402	2,20
12	6 093	2,40	7 266	2,86	24	5 481	2,41	5 591	2,47
Średnia ze wszystkich godzin						6 092	2,49	6 703	2,73

Źródło: opracowanie własne.

Przypomnijmy, że do prognozy w modelu (3.5.1) dla każdej godziny doby wykorzystaliśmy odrębną sieć neuronową MLP, przy czym modele te mają zastosowanie do dni typowych (bez świąt i dni poświątecznych – patrz punkt 2.2.5). Obecnie interesować nas będzie ocena różnic w działaniu otrzymanych sieci neuronowych, przy dokładnych i szacowanych wartościach wejściowych zmiennych temperaturowych $TMIN_d$, $TMAX_d$ (Bartkiewicz 2000e; Bartkiewicz

2001b). Po przetestowaniu modelu dla rocznego zbioru danych otrzymane błędy prognozy zebrane zostały w tabeli 3.5.1.

Drugie zagadnienie, które przeanalizujemy w bieżącym podrozdziale, związane będzie z zastosowaniem sieci neuronowej do prognozy maksymalnego godzinnego zapotrzebowania na energię, z dwudniowym wyprzedzeniem czasowym (ten sam problem, co w poprzednim podrozdziale). Model zdefiniowany jest przez równanie postaci:

$$ZS_d = f(ZS_{d-7}, ZG(t-2)_{d-2}, ZG(t-1)_{d-2}, ZS(t)_{d-2}, ZG(t+1)_{d-2}, TMIN_d, TMAX_d, dt_{1d}, ..., dt_{6d}) \quad (3.5.2)$$

gdzie:

$ZS(t)_d$ – zapotrzebowanie na energię w szczytowej godzinie, w dniu d,

$ZG(t)_d$ – zapotrzebowanie na energię o godzinie t w dniu d, gdzie t oznacza godzinę szczytową,

$TMIN_d$, $TMAX_d$ – prognozy temperatur na dzień d,

dt_{id}, $i = 1, \ldots, 6$ – zmienne kodujące dzień tygodnia.

Porównanie dokładności predykcji szczytowego zapotrzebowania na energię elektryczną, otrzymanych przy dokładnych i szacowanych wartościach wejściowych zmiennych temperaturowych $TMIN_d$, $TMAX_d$ (Bartkiewicz 2000c), przedstawione zostało w tabeli 3.5.2.

Tabela 3.5.2. Porównanie dokładności prognozy maksymalnego godzinnego zapotrzebowania na energię dla dokładnych i zaszumionych wartości temperatur

Rodzaj wartości temperatur	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
Temperatury dokładne	7 918	29 550	2,88	16,75
Prognozy temperatur	8 523	30 338	3,07	17,19

Źródło: opracowanie własne.

Analizując wyniki przedstawione w tabelach 3.5.1 oraz 3.5.2, możemy powiedzieć, że błąd prognoz temperatur nie ma jakiegoś ogromnego wpływu na dokładność prognozy zapotrzebowania na energię, ale jednak należy go uznać za dosyć wyraźny. Różnica rzędu 0,3% powoduje, że musimy przyjrzeć się temu zagadnieniu bliżej i zastanowić się nad jego uwzględnieniem podczas modelowania niepewności rozkładu prognozowanego zapotrzebowania na energię.

W dalszej części bieżącego podrozdziału zajmujemy się więc problematyką wnioskowania na temat zmienności wyjścia modelu (prognozy) związanej z niepewnością wejść, dla predyktorów w postaci jednokierunkowych sieci

neuronowych i neuronowo-rozmytych. Rozważania te łatwo mogą być jednak przełożone na ogólny przypadek dowolnego nieliniowego modelu prognostycznego. Skupimy się przy tym na problemie predykcji, pomijając całkowicie zagadnienia uczenia modelu w warunkach występowania szumu dla zmiennych wejściowych.

Musimy najpierw zastanowić się nad przyczynami pogorszenia dokładności działania naszego modelu prognostycznego. Przyjmijmy więc, podobnie jak w punkcie 3.1.1, że mamy do dyspozycji wytrenowaną sieć neuronową lub neuronowo-rozmytą $f(\mathbf{x}, \mathbf{w})$ o charakterze regresyjnym, tzn. realizującą pewne odwzorowanie stochastyczne między ciągłą zmienną objaśnianą y, czyli w naszym przypadku zapotrzebowaniem na energię elektryczną, oraz zmienną objaśniającą $\mathbf{x}$:

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon \tag{3.5.3}$$

gdzie $\mathbf{w}$ jest zbiorem parametrów (wag) sieci, zaś ε addytywnym czynnikiem losowym. Przyjmujemy również, że nasza zmienna objaśniana y to zmienna skalarna, natomiast $\mathbf{x} = (x_1, \ldots, x_n)$ jest zmienną wektorową złożoną z n zmiennych wejściowych. Zakładamy dalej, że model został nauczony prawidłowo i jest nieobciążonym estymatorem wartości oczekiwanej w rozkładzie warunkowym zmiennych $\mathbf{x}$ i y, tzn. tak jak to pokazaliśmy w punkcie 3.1.1, możemy przyjąć, że:

$$E(y/\mathbf{x}) = \int_{-\infty}^{\infty} yp(y/\mathbf{x})dy \approx f(\mathbf{x}, \mathbf{w}) \tag{3.5.4}$$

Oszacowanie parametrów modelu nastąpiło poprzez minimalizację błędu kwadratowego na zbiorze treningowym $D = \{\mathbf{x}_k, y_k\}$, $k = 1, \ldots, N$, dla którego wartości wykorzystywanych wzorców danych znane były dokładnie.

Wykorzystując nasz model do wyznaczenia prognozy, wartości niektórych wejść jesteśmy jednak w stanie określić jedynie w przybliżeniu, w postaci prognoz lub oszacowań. Wyodrębnijmy ze zbioru zmiennych wejściowych x_i, $i = 1, \ldots, n$ te wejścia, których wartości znane są dokładnie i oznaczmy je przez $\mathbf{x}^d = (x_1^d, \ldots, x_{zn}^d)$. Pozostałe zmienne wejściowe, czyli w naszym przypadku obarczone błędem zmienne temperaturowe, oznaczmy przez $\mathbf{x}^z = (x_1^z, \ldots, x_{sz}^z)$, gdzie oczywiście $zn + sz = n$. Zakładamy, że dane są w postaci obserwowanych oszacowań (prognoz) $\mathbf{z} = (z_1, \ldots, z_{sz})$, związanych z wartościami $\mathbf{x}^z$ zależnością:

$$\mathbf{x}^z = \mathbf{z} + \boldsymbol{\delta} \tag{3.5.5}$$

gdzie $\boldsymbol{\delta}$ jest addytywnym szumem losowym wynikającym z błędu oszacowań powyższych zmiennych wejściowych. Zakładać ponadto będziemy, że $\mathbf{z}$ jest nieobciążonym oszacowaniem prawdziwych wartości zmiennych wejściowych $\mathbf{x}^z$, czyli wartość oczekiwana błędów $\boldsymbol{\delta}$ jest równa zero.

Pierwszym istotnym źródłem spadku dokładności prognozy może być fakt, że nasza sieć neuronowa (neuronowo-rozmyta) stanowi przybliżenie wartości oczekiwanej prognozowanego zapotrzebowania na energię względem dokładnych wartości temperatur, a nie ich oszacowań, które wykorzystujemy na wejściu prognozy. Wobec tego powstaje pytanie, czy i jak możemy wywnioskować funkcję regresji zmiennej y względem obserwowanych wartości zmiennej $\mathbf{z}$, zamiast nieznanych wartości zmiennej $\mathbf{x}^z$ (oczywiście dla danej konkretnej wartości znanych wejść $\mathbf{x}^d$). Wartość oczekiwana $E(y / \mathbf{z}, \mathbf{x}^d)$ dana jest, z definicji, wzorem:

$$E(y / \mathbf{z}, \mathbf{x}^d) = \int_{-\infty}^{\infty} y p(y / \mathbf{z}, \mathbf{x}^d) dy \tag{3.5.6}$$

Korzystając ze wzoru na prawdopodobieństwo całkowite dla funkcji gęstości ciągłych rozkładów prawdopodobieństwa:

$$p(y) = \int_{-\infty}^{\infty} p(y / t) p(t) dt \tag{3.5.7}$$

możemy rozwinąć wzór na prawdopodobieństwo warunkowe w (3.5.6) względem nieznanej zmiennej prawdziwej wartości zmiennej $\mathbf{x}^z$:

$$\begin{aligned} p(y / \mathbf{z}, \mathbf{x}^d) &= \int_{-\infty}^{\infty} p(y / \mathbf{z}, \mathbf{x}^z, \mathbf{x}^d) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z = \\ &= \int_{-\infty}^{\infty} p(y / \mathbf{z}, \mathbf{x}) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z \end{aligned} \tag{3.5.8}$$

gdzie $p(\mathbf{x}^z / \mathbf{z})$ jest gęstością rozkładu prawdopodobieństwa wartości zmiennych $\mathbf{x}^z$, przy danych ich oszacowaniach $\mathbf{z}$ (czyli błędu oszacowania).

Ponieważ jednak prawdopodobieństwo w rozkładzie warunkowym $y / \mathbf{x}$ nie zależy w żaden sposób od $\mathbf{z}$ (zakładamy bowiem, że model budowany jest dla dokładnych wartości $\mathbf{x}^z$, a nie dla ich prognoz $\mathbf{z}$, czyli zmienna y zależna jest od $\mathbf{z}$ tylko za pośrednictwem $\mathbf{x}^z$), to (3.5.8) możemy przepisać jako:

$$p(y/\mathbf{z},\mathbf{x}^d)=\int_{-\infty}^{\infty}p(y/\mathbf{x})p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z \tag{3.5.9}$$

Podstawiając wyznaczoną zależność (3.5.9) dla funkcji gęstości prawdopodobieństwa warunkowego rozkładu zmiennej wyjściowej do początkowej zależności (3.5.6) na wartość oczekiwaną prognozy względem obserwowanych (zaszumionych) wartości wejściowych $E(y / \mathbf{z}, \mathbf{x}^d)$, otrzymujemy:

$$E(y/\mathbf{z},\mathbf{x}^d)=\int_{-\infty}^{\infty}y\left(\int_{-\infty}^{\infty}p(y/\mathbf{x})p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z\right)dy=$$

$$=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}yp(y/\mathbf{x})dy\right)p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z= \tag{3.5.10}$$

$$=\int_{-\infty}^{\infty}E(y/\mathbf{x})p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z$$

Ponieważ wartość oczekiwana $E(y / \mathbf{x})$ określona jest za pomocą sieci neuronowej (neuronowo-rozmytej), ostatecznie więc możemy nieobciążony estymator funkcji regresji y względem $(\mathbf{z}, \mathbf{x}^d)$ zapisać jako:

$$E(y/\mathbf{z},\mathbf{x}^d)=\int_{-\infty}^{\infty}f(\mathbf{x}^z,\mathbf{x}^d,\mathbf{w})p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z \tag{3.5.11}$$

Stosując identyczne rozumowanie, dodatkową wariancję prognozy (wyjścia modelu) spowodowaną niepewnością jego wejść $\sigma^2_{\mathbf{x}}(\mathbf{x})$ możemy wyznaczyć przy użyciu następującej zależności:

$$\sigma^2_{\mathbf{x}}(\mathbf{x})=\int_{-\infty}^{\infty}(f(\mathbf{z},\mathbf{x}^d,\mathbf{w})-f(\mathbf{x}^z,\mathbf{x}^d,\mathbf{w}))^2\,p(\mathbf{x}^z/z)d\mathbf{x}^z \tag{3.5.12}$$

Wartość oczekiwaną zapotrzebowania na energię elektryczną w rozkładzie warunkowym względem obserwowanych wartości wejść modelu prognostycznego (zarówno znanych dokładnie, jak i szacowanych), wyznaczyć więc można, całkując wyjścia sieci neuronowej (neuronowo-rozmytej) względem rozkładu prawdopodobieństwa prognozowanych wartości wejść $\mathbf{z}$. W podobny sposób da się określić dodatkową wariancję prognozowanego zapotrzebowania, spowodo-

waną niepewnością wykorzystywanych danych wejściowych dotyczących temperatur.

Problem polega na tym, że całki we wzorach (3.5.11) oraz (3.5.12) są skomplikowanymi całkami wielowymiarowymi, których rozwiązania uzyskać można numerycznie, stosując różnego rodzaju warianty metod Monte Carlo (Tresp, Hofman 1998; Tresp, Neuneier, Ahmad 1995; Wright 1999). Koszty obliczeniowe tych metod mogą być jednak bardzo wysokie, w zależności od liczby zaszumionych zmiennych wejściowych *sz*. Dlatego zajmiemy się najpierw metodami uproszczonymi pozwalającymi uzyskać przybliżone wartości (3.5.11) i (3.5.12).

3.5.2. Oszacowania oparte na lokalnej linearyzacji modelu

W bieżącym punkcie przeanalizujemy zagadnienie wyznaczenia wartości oczekiwanej prognozy zapotrzebowania na energię elektryczną przy wykorzystaniu obarczonych błędem wartości wejściowych (3.5.11) oraz związanej z tym dodatkowej niepewności predykcji określonej przez zależność (3.5.12), przy założeniu, że model można z dostatecznie dużą dokładnością przybliżyć za pomocą zależności liniowej.

Jeśli błędy szacowanych zmiennych wejściowych modelu prognostycznego $\boldsymbol{\delta}$ nie są zbyt duże oraz spełnione zostaną założenia twierdzenia Taylora, to znaczy zależność wyjścia modelu względem zaszumionych zmiennych nie wykazuje w rozważanym lokalnym otoczeniu $\mathbf{z}$ zbyt silnych cech nieliniowych, to możemy zastosować aproksymację pierwszego rzędu wyjścia modelu predyktora $f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w})$, rozwijając go w szereg z dokładnością do wyrazów liniowych:

$$\begin{aligned} f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) &= f(\mathbf{z} - \boldsymbol{\delta}, \mathbf{x}^d, \mathbf{w}) \approx \\ &\approx f(\mathbf{z}, \mathbf{x}^d, \mathbf{w}) - \left[\left. \frac{\partial f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w})}{\partial \mathbf{x}^z} \right|_{\mathbf{x}^z = z} \right]^{\mathrm{T}} \boldsymbol{\delta} = \\ &= f(\mathbf{z}, \mathbf{x}^d, \mathbf{w}) - \mathbf{g}\mathbf{x}_z^{\mathrm{T}}(\mathbf{z}, \mathbf{x}^d, \mathbf{w})\boldsymbol{\delta} \end{aligned} \tag{3.5.13}$$

gdzie $\mathbf{g}\mathbf{x}_z(\mathbf{z}, \mathbf{x}^d, \mathbf{w})$ jest *nz*-wymiarowym gradientem wyjścia sieci neuronowej lub neuronowo-rozmytej wyznaczonym względem jej zaszumionych zmiennych wejściowych $\mathbf{x}^z$, dla wartości $\mathbf{x}^z = \mathbf{z}$.

Podstawiając teraz przybliżoną postać liniową równania sieci neuronowej (neuronowo-rozmytej) (3.5.13) do wyznaczonej w poprzednim punkcie ogólnej formuły dla wartości oczekiwanej prognozy względem obciążonych błędem

wartości wejściowych modelu, określonej przez zależność (3.5.11), otrzymujemy:

$$
\begin{aligned}
E(y/\mathbf{z},\mathbf{x}^d) &= \int_{-\infty}^{\infty} f(\mathbf{x}^z,\mathbf{x}^d,\mathbf{w})\,p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z \approx \\
&\approx \int_{-\infty}^{\infty} (f(\mathbf{z},\mathbf{x}^d,\mathbf{w}) - \mathbf{g}\mathbf{x}_z^T(\mathbf{z},\mathbf{x}^d,\mathbf{w})\boldsymbol{\delta})\,p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z = \\
&= \int_{-\infty}^{\infty} f(\mathbf{z},\mathbf{x}^d,\mathbf{w})\,p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z - \int_{-\infty}^{\infty} \mathbf{g}\mathbf{x}_z^T(\mathbf{z},\mathbf{x}^d,\mathbf{w})\boldsymbol{\delta}p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z = \\
&= f(\mathbf{z},\mathbf{x}^d,\mathbf{w}) \int_{-\infty}^{\infty} p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z - \sum_{i=1}^{sz} gx_{x_i^z}(z_i,\mathbf{x}^d,\mathbf{w}) \int_{-\infty}^{\infty} \delta_i p(x_i^z/z_i)dx_i^z
\end{aligned}
\tag{3.5.14}
$$

gdzie sumowanie w drugim członie (3.5.14) przebiega po wszystkich zmiennych wejściowych modelu obarczonych błędem.

Zauważmy, że całka występująca w pierwszym członie (3.5.14), jako całka z gęstości rozkładu w całej przestrzeni, jest równa 1:

$$
f(\mathbf{z},\mathbf{x}^d,\mathbf{w}) \int_{-\infty}^{\infty} p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z = f(\mathbf{z},\mathbf{x}^d,\mathbf{w}) \tag{3.5.15}
$$

Ponadto zakładaliśmy, że $\mathbf{z}$ jest nieobciążonym oszacowaniem prawdziwych wartości zmiennych wejściowych $\mathbf{x}^z$, czyli wartości oczekiwane błędów δ_i są równe zero. Całki występujące w drugim członie (3.5.14) zanikają więc, zaś wartości pochodnych modelu względem zaszumionych wejść, obliczane w punkcie z_i, są stałymi. Możemy zatem zapisać:

$$
\sum_{i=1}^{sz} gx_{x_i^z}(z_i,\mathbf{x}^d,\mathbf{w}) \int_{-\infty}^{\infty} \delta_i p(x_i^z/z_i)dx_i^z = \sum_{i=1}^{sz} gx_{x_i^z}(z_i,\mathbf{x}^d,\mathbf{w}) \cdot 0 = 0 \tag{3.5.16}
$$

Ostatecznie więc, podstawiając (3.5.15) oraz (3.5.16) do (3.5.14), wartość oczekiwaną prognozy zapotrzebowania na energię, względem obciążonych błędem danych wejściowych modelu, możemy zapisać jako:

$$
E(y/\mathbf{z},\mathbf{x}^d) \approx f(\mathbf{z},\mathbf{x}^d,\mathbf{w}) \tag{3.5.17}
$$

Innymi słowy, optymalna strategia postępowania polega w tym przypadku na zastąpieniu nieznanych dokładnych wartości wejść $\mathbf{x}^z$ ich znanymi oszacowaniami $\mathbf{z}$. W praktyce jest to oczywiście powszechny i najczęściej stosowany sposób działania ze zmiennymi wejściowymi o zaszumionym charakterze. Zwróćmy jednak uwagę, że u podstaw jego zastosowania leży lokalna aproksymacja modelu za pomocą funkcji liniowej (3.5.13). Jeżeli założenia wzoru Taylora spełnione będą jedynie w przybliżeniu, z powodu zbyt dużych błędów $\boldsymbol{\delta}$ albo znaczącego wpływu nieliniowego charakteru zależności wyjścia sieci neuronowej (neuronowo-rozmytej) od zaszumionych zmiennych wejściowych, co wyraża się w dużych wartościach pochodnych wyższych rzędów rozwinięcia funkcji w szereg, to wówczas oszacowanie (3.5.13) może być obarczone dużym błędem. Oznaczałoby to, że przybliżenie wartości oczekiwanej dane przez (3.5.17) może stać się mało dokładne, skutkując obciążeniem tego estymatora.

Jeśli z kolei założymy, że przybliżenie wartości oczekiwanej zmiennej wyjściowej (3.5.17) jest dostatecznie dobre, a otrzymywana prognoza – nieobciążona, to głównym źródłem przyrostu błędu będzie dodatkowa wariancja prognozy $\sigma^2_{\mathbf{x}}(\mathbf{x})$, spowodowana niepewnością danych wejściowych (propagacją ich błędów przez model). Przypomnijmy, że wariancję spowodowaną przez użycie dla sieci neuronowej lub neuronowo-rozmytej przybliżeń $\mathbf{z}$, zamiast dokładnych wartości wejściowych $\mathbf{x}^z$, możemy wyznaczyć przy użyciu zależności (3.5.12). Spróbujmy teraz oszacować tę formułę, wykorzystując podejście oparte na lokalnej linearyzacji modelu prognostycznego.

Ponownie rozwijając funkcję sieci neuronowej (neuronowo-rozmytej) $f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w})$ w szereg Taylora, w otoczeniu oszacowanych wartości $\mathbf{z}$, otrzymujemy:

$$\begin{aligned}\sigma^2_{\mathbf{x}}(\mathbf{x}) &= \int_{-\infty}^{\infty} (f(\mathbf{z}, \mathbf{x}^d, \mathbf{w}) - f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}))^2 \, p(\mathbf{x}^z / z) d\mathbf{x}^z \approx \\ &\approx \int_{-\infty}^{\infty} (f(\mathbf{z}, \mathbf{x}^d, \mathbf{w}) - f(\mathbf{z}, \mathbf{x}^d, \mathbf{w}) + \mathbf{g}\mathbf{x}_z^T(\mathbf{z}, \mathbf{x}^d, \mathbf{w})\boldsymbol{\delta})^2 \, p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z = \\ &= \int_{-\infty}^{\infty} (\mathbf{g}\mathbf{x}_z^T(\mathbf{z}, \mathbf{x}^d, \mathbf{w})\boldsymbol{\delta})^2 \, p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z\end{aligned} \tag{3.5.18}$$

gdzie $\mathbf{g}\mathbf{x}_z(\mathbf{z}, \mathbf{x}^d, \mathbf{w})$, podobnie jak w (3.5.13), oznacza nz-wymiarowy gradient wyjścia sieci neuronowej lub neuronowo-rozmytej, wyznaczony względem jej zaszumionych zmiennych wejściowych $\mathbf{x}^z$, dla wartości $\mathbf{x}^z = \mathbf{z}$, zaś $\boldsymbol{\delta}$ jest błędem oszacowań wartości tych zmiennych.

Korzystając teraz w (3.5.18) z definicji iloczynu skalarnego, otrzymujemy:

$$\sigma_{\mathbf{x}}^2(\mathbf{x}) \approx \int_{-\infty}^{\infty} (\mathbf{g}\mathbf{x}_z^T(\mathbf{z},\mathbf{x}^d,\mathbf{w})\boldsymbol{\delta})^2 p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z =$$

$$= \int_{-\infty}^{\infty} \left(\sum_{i=1}^{sz} gx_{x_i^z}(z_i,\mathbf{x}^d,\mathbf{w})\delta_i \right) \left(\sum_{j=1}^{sz} gx_{x_j^z}(z_j,\mathbf{x}^d,\mathbf{w})\delta_j \right) p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z = \qquad (3.5.19)$$

$$= \int_{-\infty}^{\infty} \left(\sum_{i=1}^{sz}\sum_{j=1}^{sz} gx_{x_i^z}(z_i,\mathbf{x}^d,\mathbf{w})\delta_i gx_{x_j^z}(z_j,\mathbf{x}^d,\mathbf{w})\delta_j \right) p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z$$

Pogrupujmy w (3.5.19) odrębnie składniki z gradientami oraz z błędami poszczególnych zmiennych wejściowych δ. Zauważmy ponadto, że wykorzystywane gradienty względem wejść modelu oblicza się w punkcie $\mathbf{z}$, są więc one stałe ze względu na $\mathbf{x}^z$. Możemy zatem wyciągnąć je spod znaku całki. I tak otrzymujemy:

$$\sigma_{\mathbf{x}}^2(\mathbf{x}) \approx \sum_{i=1}^{sz}\sum_{j=1}^{sz} gx_{x_i^z}(z_i,\mathbf{x}^d,\mathbf{w})gx_{x_j^z}(z_j,\mathbf{x}^d,\mathbf{w}) \int_{-\infty}^{\infty} \delta_i\delta_j p(x_i^z,x_j^z/\mathbf{z})dx_i^z dx_j^z \qquad (3.5.20)$$

Przypomnijmy, że kowariancja błędów δ_i, δ_j wartości zmiennych wejściowych x_i^z oraz x_j^z z definicji jest równa:

$$\mathrm{cov}(\delta_i,\delta_j) = \int_{-\infty}^{\infty} \delta_i\delta_j p(x_i^z,x_j^z/\mathbf{z})dx_i^z dx_j^z \qquad (3.5.21)$$

Zauważmy więc, że całka występująca w (3.5.20) równa jest kowariancji błędów wartości zmiennych wejściowych δ_i, δ_j. Jeżeli teraz oznaczymy przez $\mathbf{C}_{\boldsymbol{\delta}}$ macierz kowariancji błędów zaszumionych zmiennych x_i^z oraz x_j^z, to oszacowanie (3.5.20) dodatkowej wariancji prognozy $\sigma_{\mathbf{x}}^2(\mathbf{x})$, spowodowanej niepewnością danych wejściowych, możemy ostatecznie zapisać w postaci:

$$\sigma_{\mathbf{x}}^2(\mathbf{x}) \approx \mathbf{g}\mathbf{x}_z^T(\mathbf{z},\mathbf{x}^d,\mathbf{w})\,\mathbf{C}_{\boldsymbol{\delta}}\,\mathbf{g}\mathbf{x}_z(\mathbf{z},\mathbf{x}^d,\mathbf{w}) \qquad (3.5.22)$$

Otrzymany wynik nie wydaje się specjalnie zaskakujący. Mamy tu oczywiście do czynienia ze znanym z większości tekstów statystycznych o średnim poziomie zaawansowania prawem propagacji błędów – w szerszej wersji uwzględniającej błędy skorelowane. Zwróćmy jednak ponownie uwagę, że

oszacowanie to jest poprawne, o ile poprawna jest linearyzacja modelu poprzez rozwinięcie go w szereg Taylora.

Zauważmy ponadto, że do wyznaczenia dodatkowej wariancji prognozy, wynikającej z niepewności wejść, niezbędne jest wyznaczenie gradientu wyjścia modelu względem zaszumionych zmiennych wejściowych $\mathbf{gx}_z(\mathbf{z}, \mathbf{x}^d, \mathbf{w})$. W ogólnym przypadku wyznaczony może być on numerycznie, jednakże dla sieci neuronowych lub neuronowo-rozmytych rozważanych w tej pracy dostępne są odpowiednie procedury analityczne. Dla warstwowych sieci perceptronowych MLP problem ten omawiamy w załączniku 1 w punkcie Z1.4. Podobnie w przypadku sieci z funkcjami o bazie rozmytej FBF, algorytm wyznaczania gradientu wyjścia względem wejść prezentowany jest w załączniku 2 w punkcie Z2.4, zaś dla sieci neuronowo-rozmytych typu Takagi–Sugeno, z liniowymi funkcjami następników reguł – w załączniku 3 w punkcie Z3.4.

Przedstawione podejście przetestowaliśmy dla zestawu sieci neuronowych MLP wykorzystanych do prognozy godzinnego zapotrzebowania na energię elektryczną, z dwudniowym wyprzedzeniem czasowym (model (3.5.1)). Obecnie więc zakładać będziemy, że wariancja rozkładu prawdopodobieństwa prognozowanego zapotrzebowania na energię $\sigma^2_y(\mathbf{x})$, dla danego wzorca danych wejściowych $\mathbf{x}$, szacowana jest za pomocą trzech niezależnych komponentów związanych z różnymi źródłami niepewności:

$$\sigma^2_y(\mathbf{x}) = \sigma^2_{\mathbf{w}}(\mathbf{x}) + \sigma^2_{\varepsilon}(\mathbf{x}) + \sigma^2_{\mathbf{x}}(\mathbf{x}) \tag{3.5.23}$$

gdzie $\sigma^2_{\mathbf{w}}(\mathbf{x})$ oznacza wariancję wyjścia modelu prognostycznego wynikającą z niepewności parametrów (wag), $\sigma^2_{\varepsilon}(\mathbf{x})$ – wariancję czynnika losowego (szum losowy), zaś $\sigma^2_{\mathbf{x}}(\mathbf{x})$ – wariancję wyjścia modelu spowodowaną niepewnością danych wejściowych. Pamiętajmy jeszcze, że do tego, by rozkład otrzymywanej prognozy miał charakter rozkładu normalnego, błąd zmiennych wejściowych modelu musi mieć rozkład normalny $N(0, \mathbf{C}_{\boldsymbol{\delta}})$.

W przeprowadzonych przez nas badaniach pierwszy komponent w (3.5.23) $\sigma^2_{\mathbf{w}}(\mathbf{x})$ wyznaczony został przy użyciu (omawianej w punkcie 3.3.2) metody delta, w wersji z dokładnym oszacowaniem hesjanu błędu. Do wyznaczenia odchylenia standardowego elementu losowego $\sigma^2_{\varepsilon}(\mathbf{x})$ wykorzystana została dodatkowa sieć neuronowa, tak jak to wskazywaliśmy w punkcie 3.4.4. Natomiast wariancja prognozy spowodowana niepewnością jej danych wejściowych $\sigma^2_{\mathbf{x}}(\mathbf{x})$ została oszacowana za pomocą (3.5.23). Odchylenie standardowe prognozy zapotrzebowania na energię elektryczną $\sigma_y(\mathbf{x})$ ostatecznie określone więc będzie w naszym przypadku zależnością:

$$\sigma_y(\mathbf{x}) = \sqrt{\sigma^2_{\varepsilon}(\mathbf{x})(1 + \mathbf{g}(\mathbf{x}, \mathbf{w}^*)^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{g}(\mathbf{x}, \mathbf{w}^*)) + \sigma^2_{\mathbf{x}}(\mathbf{x})} \tag{3.5.24}$$

Przedstawione oszacowanie odchylenia standardowego prognozy wykorzystano do wyznaczenia jej przedziałów. Przypomnijmy, że przy przyjętym poziomie prawdopodobieństwa α dolny i górny kraniec przedziału prognozy wyznaczyć możemy zgodnie z formułą:

$$\begin{aligned} d_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) - Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_y(\mathbf{x}) \\ g_{y/\mathbf{x}}(\alpha) &= f(\mathbf{x},\mathbf{w}) + Q_{N(0,1)}((1+\alpha)/2)\cdot\sigma_y(\mathbf{x}) \end{aligned} \tag{3.5.25}$$

gdzie $Q_{N(0,1)}(\alpha)$ jest kwantylem rozkładu normalnego standardowego $N(0, 1)$ (przy mniejszych rozmiarach wykorzystywanych zbiorów treningowych powinniśmy skorzystać z kwantyli rozkładu t-Studenta). Przypomnijmy ponadto, że aby oszacowanie (3.5.25) dawało poprawne wyniki, rozkład błędu zmiennych wejściowych modelu choćby w przybliżeniu odpowiadać musi rozkładowi normalnemu $N(0, \mathbf{C}_{\boldsymbol{\delta}})$ o wartości oczekiwanej zero i pewnej macierzy kowariancji $\mathbf{C}_{\boldsymbol{\delta}}$.

Do wyznaczenia na podstawie (3.5.23) wariancji prognozy, spowodowanej niepewnością zmiennych wejściowych $\sigma^2_{\mathbf{x}}(\mathbf{x})$, niezbędne jest określenie macierzy kowariancji ich błędów, $\mathbf{C}_{\boldsymbol{\delta}}$. Jak już wspomnieliśmy w punkcie 3.5.1, w modelu (3.5.1) (podobnie jak w przypadku większości rozważanych przez nas zagadnień krótkoterminowej prognozy zapotrzebowania na energię) posługujemy się prognozami temperatury minimalnej i maksymalnej. Mamy więc dwie zmienne wejściowe obarczone niepewnością (czyli w tym przypadku $sz = 2$).

Przeprowadzona analiza dostępnych prognoz temperatur minimalnych i maksymalnych oraz faktycznych wartości tych zmiennych w pewnym okresie czasu pozwoliła na empiryczne wyznaczenie macierzy kowariancji błędów wykorzystywanych prognoz. Rozkład błędu tych zmiennych wejściowych miał charakter normalny, zaś oszacowana macierz $\mathbf{C}_{\boldsymbol{\delta}}$ wyniosła:

$$\mathbf{C}_{\boldsymbol{\delta}} = \begin{bmatrix} 3{,}78 & 1{,}53 \\ 1{,}53 & 3{,}51 \end{bmatrix} \tag{3.5.26}$$

Wyniki testów przedziałów prognozy zapotrzebowania na energię elektryczną dla wybranych godzin, wyznaczonych za pomocą zależności (3.5.25), znajdują się w tabeli 3.5.3. Jak widać, badania liczby prognoz w obrębie przedziałów (częstości) dają wyniki zbliżone do teoretycznych poziomów prawdopodobieństwa, dla których je obliczono. Przedstawione testy przeprowadzono dla dużego, dwuletniego zbioru danych (z wyłączeniem nietypowych dni świątecznych i poświątecznych – patrz punkt 2.2.5).

Tabela 3.5.3. Częstości empiryczne przedziałów prognozy zapotrzebowania na energię dla wybranych godzin, z uwzględnieniem niepewności temperatur (w %)

Prawdopodobieństwo	Godz. 7	Godz. 12	Godz. 17	Godz. 20
80	81,90	80,10	80,40	83,10
85	83,70	85,20	85,50	83,40
90	92,30	88,10	91,40	87,80
95	97,00	95,80	94,40	95,50

Źródło: opracowanie własne.

3.5.3. Wyznaczanie prognozy w warunkach niepewności wejść przy użyciu metod opartych na próbkowaniu Monte Carlo

Powszechnie stosowana podczas sporządzania prognozy strategia zastępowania nieznanych prawdziwych wartości zmiennych wejściowych szacowanymi wartościami oczekiwanymi wynika z przyjętego założenia o możliwości lokalnej aproksymacji odwzorowania nieliniowego modelu sieci neuronowej (neuronowo-rozmytej) za pomocą funkcji liniowej. Pokazaliśmy to w poprzednim punkcie, wyprowadzając zależność (3.5.17). Intuicyjnie zresztą taki sposób postępowania wydaje się logiczny i oczywisty. Wiemy przecież z podstawowych kursów statystyki, że wartość oczekiwana stanowi operację liniową, czyli dla dowolnej zmiennej losowej X możemy zapisać:

$$E(aX+b) = aE(X)+b \tag{3.5.27}$$

Jeśli więc liniowe przybliżenie zależności między wyjściem sieci a zaszumionymi zmiennymi $\mathbf{x}^z$ (rozwinięcie w szereg Taylora) jest dobre w otoczeniu oszacowanych wartości tych zmiennych $\mathbf{z}$, to przybliżenie dokładnej wartości oczekiwanej prognozy zapotrzebowania na energię (3.5.11) za pomocą (3.5.17) również będzie dobre. Mówiąc prościej, w naszym przypadku podanie na wejściu modelu prognoz (wartości oczekiwanych) temperatur nie powinno powodować zwiększenia obciążenia modelu.

Przypomnijmy jednak, że zgodnie z podawanymi w punkcie 3.5.1 wynikami badań dokładności prognoz zapotrzebowania na energię, uzyskanych przy użyciu tej strategii (tabele 3.5.1 i 3.5.2), błąd predykcji w niewielkim stopniu, ale jednak widocznie, wzrastał. Powstaje wobec tego pytanie, czy wzrost błędu przewidywanego zapotrzebowania spowodowany jest wyłącznie dodatkową wariancją modelu wynikającą z propagacji niepewności prognoz temperatur, którą oszacowaliśmy w poprzednim punkcie za pomocą zależności (3.5.22). Czy też może zależność wyjścia modelu od zmiennych temperaturowych ma

charakter silnie nieliniowy i lokalna linearyzacja odwzorowania sieci neuronowej (neuronowo-rozmytej) przy użyciu szeregu Taylora jest niedokładna? Pamiętajmy bowiem, że w przypadku funkcji nieliniowej transformacja wartości oczekiwanej zmiennej wejściowej niekoniecznie musi dawać w wyniku wartość oczekiwaną zmiennej wyjściowej. Użycie na wejściu sieci prognoz (wartości oczekiwanych) temperatur niekoniecznie zatem musi dawać w wyniku wartość oczekiwaną zapotrzebowania na energię.

Gdybyśmy mieli do czynienia z tym drugim przypadkiem, istniałaby pewna szansa na redukcję błędu prognozy zapotrzebowania na energię i „odzyskanie" części dokładności utraconej z powodu niepewności prognoz temperatur. Musimy zamiast wykorzystywanego w przyjętej strategii postępowania niedokładnego oszacowania (3.5.17) znaleźć inny sposób obliczenia wartości oczekiwanej wyjścia sieci neuronowej (neuronowo-rozmytej) (3.5.11), przy określonym modelu zmiennych wejściowych. Dostępne są tutaj pewne możliwości, które możemy wykorzystać, wymagają one jednak dosyć poważnych nakładów obliczeniowych.

W związku z tym w dalszych punktach bieżącego podrozdziału spróbujemy przyjrzeć się tym możliwościom i zastanowić się nad ich wpływem na dokładność prognozy krótkoterminowego zapotrzebowania na energię. Jeszcze raz zwróćmy jednak uwagę na fakt, że zostawiamy całkowicie na boku problem treningu modelu w warunkach występowania niepewności wzorców danych wejściowych. Jest to odrębne zagadnienie, które wymagałoby zagłębienia się poważnie w teorię uczenia statystycznego, problematykę tzw. regularyzacji itp. Zakładamy więc, że sieć neuronowa lub neuronowo-rozmyta trenowana jest na danych dokładnych, natomiast w trakcie operacyjnego wykorzystania gotowego modelu do sporządzania prognoz pojawia się konieczność użycia danych zaszumionych w postaci prognoz i oszacowań. W przypadku prognoz zapotrzebowania na energię, gdzie zazwyczaj na potrzeby uczenia modelu dysponujemy dużymi zbiorami dosyć precyzyjnych danych, takie połączenie ma dużo większe znaczenie praktyczne.

Od razu również powiedzmy, że w rozważanych zagadnieniach nie spodziewamy się jakiejś znaczącej poprawy. W poprzednim punkcie przeprowadziliśmy przecież badania oszacowania wariancji prognozy zapotrzebowania na energię $\sigma^2_{\mathbf{x}}(\mathbf{x})$, wynikającej z błędów predykcji temperatury maksymalnej i minimalnej (zależność (3.5.22)). Testowanie uzyskanych na podstawie tej formuły przedziałów prognozy wskazuje na jej poprawny charakter (patrz wyniki przedstawione w tabeli 3.5.3). Przybliżenie odwzorowania modelu względem zmiennych temperaturowych przy użyciu lokalnej linearyzacji sieci wydaje się więc dostatecznie dokładne. Uważny Czytelnik powinien jednak zwrócić uwagę, że analizując na początku poprzedniego rozdziału, w punkcie 2.1.2, związek procesu zapotrzebowania na energię elektryczną i temperatur, wskazywaliśmy, że zależność ta ma charakter nieliniowy (tabela 2.1.1). Spróbujmy teraz przyj-

rzeć się temu zagadnieniu nieco bliżej, tak byśmy lepiej mogli zrozumieć powód, dla którego oszacowania oparte na lokalnej linearyzacji sieci powinny raczej działać poprawnie.

Aby wyjaśnić to zjawisko, wykorzystamy wyniki badań wrażliwości dokładności prognoz dla predyktorów neuronowych MLP przedstawionych w publikacji: Bartkiewicz 1996. W wymienionej pracy przeanalizowano kilka zagadnień związanych (m.in.) z krótkoterminowym prognozowaniem zapotrzebowania na energię. Przedstawione wyniki dotyczą zarówno obciążeń sieci elektroenergetycznej, jak i gazowej (szczegóły znajdują się w legendzie do rysunku 3.5.1), ale należy nadmienić, że przebieg obu procesów jest bardzo zbliżony. Wykonane prace polegały na badaniach wrażliwości dokładności prognozy zapotrzebowania na energię w zależności od błędów wejściowych prognoz temperatur. Przeprowadzono symulację zmian dokładności predykcji (średniego błędu względnego MAPE) dla analizowanych modeli prognostycznych, przy wygenerowanych losowo zmianach wartości wejściowych temperatur, na kolejnych wzrastających co jeden stopień poziomach błędu ±1°, ±2° itd.

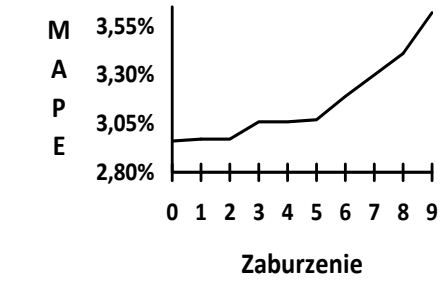
Wyniki przeprowadzonych symulacji przedstawione zostały w tabeli 3.5.4, a jednocześnie wykresy odpowiadających im krzywych błędu znajdują się na rysunku 3.5.1. Oczywiście w jednym i w drugim przypadku informacje związane z zaburzeniem wartości temperatur o 0° oznaczają działanie modelu prognostycznego dla dokładnych wartości zmiennych wejściowych.

Analizując wyniki przedstawionych symulacji, możemy stwierdzić, że modele krótkoterminowego prognozowania zapotrzebowania na energię są zasadniczo w niewielkim stopniu wrażliwe na niewielkie zaburzenia wejściowych temperatur, co może wyjaśniać również stosunkowo skromny spadek dokładności prognozy przy zastąpieniu ich dokładnych wartości oszacowaniami (tabele 3.5.1 oraz 3.5.2 w punkcie 3.5.1). Ponadto charakterystyka wrażliwości wyjścia modelu, w otoczeniu wartości dokładnej, przy niewielkich błędach temperatur jest dosyć płaska, zbliżona do liniowej. Dopiero większe błędy wejść powodują silny przyrost tempa niedokładności prognozy zapotrzebowania. Mogłoby to w pewnym sensie korespondować z naszymi obserwacjami poczynionymi w punkcie 2.2.3. Zależność między zapotrzebowaniem na energię a temperaturami ma charakter nieliniowy, ale najsilniej ujawnia się on (patrz rysunek 2.2.4) przy dużych zmianach wartości temperatur.

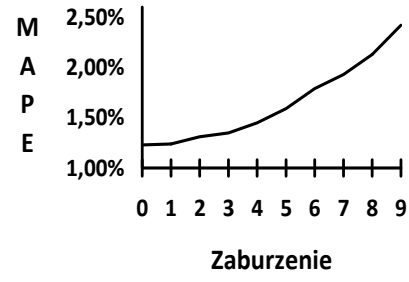
Tabela 3.5.4. Wrażliwość dokładności prognoz zapotrzebowania na energię (MAPE) na błędy wejściowych temperatur (w %)

Zaburzenie	0°	±1°	±2°	±3°	±4°	±5°	±6°	±7°	±8°	±9°
Model (a)	2,96	2,97	2,97	3,06	3,06	3,07	3,19	3,30	3,41	3,41
Model (b)	1,23	1,24	1,31	1,35	1,45	1,59	1,79	1,93	2,13	2,42
Model (c)	1,25	1,25	1,41	1,66	1,96	2,18	2,50	2,70	3,18	3,55

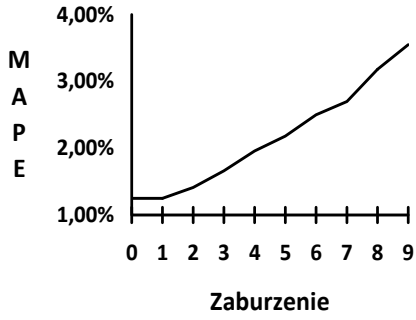
Źródło: opracowanie własne.



(a)



(b)



(c)

(a) Prognoza szczytowego zapotrzebowania na energię elektryczną
(b) Prognoza dobowego zapotrzebowania na energię elektryczną
(c) Prognoza dobowego zapotrzebowania na gaz

Rysunek 3.5.1. Wykresy krzywych ilustrujących wrażliwość dokładności prognoz zapotrzebowania na energię (MAPE) na błędy wejściowych temperatur
Źródło: opracowanie własne

Na podstawie przeprowadzonych rozważań możemy więc powiedzieć, że mamy pewne powody, aby oczekiwać tego, iż w zagadnieniach krótkoterminowej prognozy zapotrzebowania na energię lokalna linearyzacja modelu neuronowego (i podobnie neuronowo-rozmytego), w niewielkim otoczeniu dokładnych wartości wejściowych temperatur, powinna dawać w miarę poprawne wyniki. Strategia postępowania polegająca na zastąpieniu ich oszacowaniami jest więc poprawna, a przybliżenie (3.5.17) jest na tyle dokładne, że nie powoduje znaczącego obciążenia otrzymywanej prognozy. Tym niemniej pamiętać jednak należy, że problem ten zależy od konkretnego zadania prognostycznego i modelowanego systemu. W niektórych przypadkach możemy dojść do zupełnie odmiennych wniosków. Z tego powodu w dalszej części bieżącego podrozdziału przyjrzymy się wybranym metodom potencjalnie dokładniejszego wyznaczania wartości oczekiwanej prognozowanego zapotrzebowania na energię (3.5.11), choćby po to, by przekonać się, czy rzeczywiście nie będziemy w stanie uzyskać lepszych wyników.

Przypomnijmy, że $E(y / \mathbf{z}, \mathbf{x}^d)$ (czyli wartość oczekiwana prognozy zmiennej y, dla danego wzorca wejściowego złożonego ze znanych dokładnie wartości zmiennych $\mathbf{x}^d$ oraz znanych w sposób przybliżony wartości $\mathbf{z}$, zmiennych $\mathbf{x}^z$) dla danego modelu prognostycznego w formie sieci neuronowej (neuronowo-rozmytej) $f(\mathbf{x}, \mathbf{w})$ określony jest za pomocą zależności (3.5.11). Przepiszmy ją tutaj, byśmy mogli dokładniej zastanowić się nad metodami jej obliczania:

$$E(y / \mathbf{z}, \mathbf{x}^d) = \int_{-\infty}^{\infty} f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z \tag{3.5.28}$$

gdzie $p(\mathbf{x}^z / \mathbf{z})$ jest gęstością rozkładu prawdopodobieństwa wartości zmiennych $\mathbf{x}^z$, przy danych ich oszacowaniach $\mathbf{z}$ (czyli błędu oszacowania).

Problem stanowi naturalnie obliczenie całki występującej w (3.5.28). Jest to złożona wielowymiarowa całka, która musi być wyznaczana numerycznie. Choć dla całek jednowymiarowych mamy całe bogactwo różnego rodzaju kwadratur, to nie dysponujemy zbyt wieloma metodami numerycznymi obliczania całek wielowymiarowych. Spośród kilku dostępnych przeanalizujemy tutaj dwie wybrane – najczęściej stosowane w omawianych zagadnieniach.

Pierwsza metoda, jaką wykorzystamy, polega na całkowaniu z wykorzystaniem próbkowania Monte Carlo. Algorytm jest koncepcyjnie dosyć trywialny, ale daje zupełnie dobre wyniki w zadaniach związanych z całkowaniem funkcji przy danym rozkładzie prawdopodobieństwa. Biorąc więc pod uwagę ubóstwo dostępnych środków, jest to chyba najczęściej stosowane podejście w tego typu zadaniach, pomimo typowych dla metod Monte Carlo problemów efektywnościowych.

Dla przypadku ogólnego formułę całkowania pewnej funkcji $h(\mathbf{u})$ pod względem rozkładu prawdopodobieństwa $p(\mathbf{u})$ zmiennej $\mathbf{u}$, z wykorzystaniem próbkowania Monte Carlo, przedstawić możemy za pomocą następującej prostej zależności:

$$\int_{-\infty}^{\infty} h(\mathbf{u}) p(\mathbf{u}) d\mathbf{u} \approx \frac{1}{S} \sum_{i=1}^{S} h(\mathbf{u}_i) \tag{3.5.29}$$

gdzie $\mathbf{u}_1, \ldots, \mathbf{u}_S$ są próbkami losowymi pobranymi z rozkładu $p(\mathbf{u})$.

W naszym przypadku by wyznaczyć wartość oczekiwaną (3.5.28), musimy wygenerować próbę $\mathbf{u}_1, \ldots, \mathbf{u}_S$ wzorców losowych danych z rozkładu prawdopodobieństwa $p(\mathbf{x}^z / \mathbf{z})$ określającego możliwe wartości zmiennych $\mathbf{x}^z$ przy danych ich oszacowaniach $\mathbf{z}$, a następnie wykorzystać ją analogicznie jak w formule (3.5.29):

$$E(y / \mathbf{z}, \mathbf{x}^d) = \int_{-\infty}^{\infty} f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z \approx \frac{1}{S} \sum_{i=1}^{S} f(\mathbf{u}_i, \mathbf{x}^d, \mathbf{w}) \tag{3.5.30}$$

Zamiast próbkować bezpośrednio rozkład $p(\mathbf{x}^z / \mathbf{z})$ możemy, rzecz jasna, pobrać próbki z rozkładu błędu $\boldsymbol{\delta}$ oszacowania wartości zmiennych wejściowych $\mathbf{z}$. Jeżeli więc $\boldsymbol{\delta}_1$, …, $\boldsymbol{\delta}_S$ jest wygenerowaną serią próbek z rozkładu błędu $\boldsymbol{\delta}$, to pamiętając, że $\mathbf{x}^z = \mathbf{z} + \boldsymbol{\delta}$ (patrz (3.5.5)), możemy zapisać (3.5.30) jako:

$$E(y / \mathbf{z}, \mathbf{x}^d) = \int_{-\infty}^{\infty} f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z \approx \frac{1}{S} \sum_{i=1}^{S} f(\mathbf{z} + \boldsymbol{\delta}_i, \mathbf{x}^d, \mathbf{w}) \tag{3.5.31}$$

Zastosujmy więc obecnie formułę całkowania (3.5.31) do przedstawionych w punkcie 3.5.1 zagadnień krótkoterminowego prognozowania zapotrzebowania na energię elektryczną, które służą nam w bieżącym podrozdziale jako zadania testowe i weryfikujemy na nich praktycznie działanie analizowanych metod. Przypomnijmy, że określiliśmy tam dwa problemy: pierwszy dotyczący prognozy godzinnego zapotrzebowania na energię elektryczną, z dwudniowym wyprzedzeniem czasowym (zależność (3.5.1)) i drugi dotyczący prognozy maksymalnego godzinnego zapotrzebowania na energię, z dwudniowym wyprzedzeniem czasowym (zależność (3.5.2)). Jako technikę prognozowania wykorzystywaliśmy w testowanych przypadkach warstwowe sieci perceptronowe MLP.

Tabela 3.5.5. Porównanie dokładności prognoz godzinnego zapotrzebowania na energię dla dokładnych wartości temperatur i metodą próbkowania Monte Carlo

Godz.	Temperatury dokładne		Monte Carlo		Godz.	Temperatury dokładne		Monte Carlo	
	MAE (kWh)	MAPE (%)	MAE (kWh)	MAPE (%)		MAE (kWh)	MAPE (%)	MAE (kWh)	MAPE (%)
1	6 262	2,99	5 958	2,86	13	6 429	2,57	7 687	3,07
2	4 863	2,39	4 914	2,43	14	5 600	2,25	7 508	2,98
3	4 438	2,24	4 563	2,31	15	6 450	2,52	8 294	3,21
4	3 970	2,01	4 284	2,18	16	9 196	3,60	10 394	4,07
5	4 130	2,06	4 440	2,22	17	8 385	3,19	9 370	3,56
6	4 544	2,21	4 867	2,37	18	7 918	2,89	8 395	3,03
7	6 292	2,71	6 625	2,87	19	7 482	2,69	7 516	2,67
8	6 642	2,71	6 953	2,83	20	4 909	1,61	5 287	1,73
9	6 239	2,51	6 577	2,65	21	5 822	1,99	6 277	2,14
10	6 631	2,61	7 005	2,75	22	6 203	2,36	6 200	2,36
11	7 153	2,81	7 734	3,04	23	5 078	2,07	5 275	2,15
12	6 093	2,40	7 179	2,83	24	5 481	2,41	5 425	2,39
Średnia ze wszystkich godzin						6 092	2,49	6 614	2,70

Źródło: opracowanie własne.

W przypadku obydwu modeli prognostycznych zmiennymi wejściowymi obarczonymi niepewnością były temperatura maksymalna i minimalna.

W punkcie 3.5.2 stwierdziliśmy, że rozkład prawdopodobieństwa błędu oszacowań temperatur $\boldsymbol{\delta}$ ma charakter rozkładu normalnego $N(0, \mathbf{C}_{\boldsymbol{\delta}})$ o wartości oczekiwanej zero i macierzy kowariancji $\mathbf{C}_{\boldsymbol{\delta}}$, określonej przez zależność (3.5.26). Obydwie prognozy sporządzone zostały dla tego samego regionalnego systemu elektroenergetycznego i testowano je w tych samych okresach.

Po wygenerowaniu próby złożonej z serii liczb losowych o rozkładzie $N(0, \mathbf{C}_{\boldsymbol{\delta}})$ wyznaczyliśmy za pomocą formuły (3.5.31) krótkoterminowe prognozy zapotrzebowania dla obydwu modeli. Błędy predykcji w przypadku godzinnego zapotrzebowania na energię elektryczną przedstawione zostały w tabeli 3.5.5, natomiast dla maksymalnego godzinnego zapotrzebowania na energię – w tabeli 3.5.6.

Jak widzimy, zgodnie zresztą z naszymi przewidywaniami, nie udało się uzyskać znacznego postępu. Porównując znajdujące się w tabeli 3.5.5 błędy predykcji godzinnego zapotrzebowania na energię dla prognoz otrzymanych z wykorzystaniem całkowania numerycznego metodą próbkowania Monte Carlo za pomocą zależności (3.5.31) z błędami w tabeli 3.5.1 wynikającymi z użycia na wejściu modelu zaszumionych prognoz temperatur (zależność (3.5.17)), widzimy, że różnice są niewielkie. Średni błąd prognozy po scałkowaniu względem rozkładu błędów temperatur zmniejszył się zaledwie z 2,73% (tabela 3.5.1) do 2,70% (tabela 3.5.5). Podobnie na tym samym poziomie kształtują się błędy dla poszczególnych godzin.

Tabela 3.5.6. Porównanie dokładności prognoz maksymalnego godzinnego zapotrzebowania na energię dla dokładnych wartości temperatur i metodą Monte Carlo

Metoda	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
Temperatury dokładne	7 918	29 550	2,88	16,75
Monte Carlo	8 360	30 044	3,01	17,02

Źródło: opracowanie własne.

Ten sam efekt możemy zaobserwować w przypadku prognozy maksymalnego godzinnego zapotrzebowania na energię. Porównując wyniki w tabeli 3.5.2 oraz 3.5.6, obserwujemy spadek błędu z 3,07%, dla użycia na wejściu modelu temperatur zaszumionych, do 3,01%, dla prognozy, w której wykorzystano oszacowanie całki (3.5.31) metodą próbkowania Monte Carlo. Różnica w dokładności jest tym razem co prawda nieco większa, ale nadal nieznaczna.

Nie udało się nam zatem znacząco zredukować przyrostu błędów prognozy zapotrzebowania na energię, spowodowanego korzystaniem z prognoz wejściowych zmiennych temperaturowych. Tak jak przewidywaliśmy, lokalna linearyzacja modelu neuronowego lub neuronowo-rozmytego dla stosunkowo niewielkich błędów prognoz temperatur w krótkim horyzoncie czasowym nie powoduje znaczniejszego obciążenia prognozy. Przyrost błędu wynika więc przede

wszystkim ze wzrostu wariancji prognozy, a nie z niedokładności aproksymacji wartości oczekiwanej zapotrzebowania. Jeszcze raz jednak zwróćmy uwagę na fakt, że nie zawsze musi tak być. Zwłaszcza modele zapotrzebowania na energię, w których wykorzystuje się w szerszym zakresie zmienne wejściowe definiujące stan warunków atmosferycznych (takie jak ciśnienie atmosferyczne, wilgotność powietrza lub siła wiatru), powinny wykazywać większą wrażliwość na błędy tych wielkości.

Powiedzmy jeszcze kilka słów o kwestiach implementacyjnych, które pojawiają się w przypadku metody całkowania z wykorzystaniem próbkowania Monte Carlo. Formuła (3.5.31) zamienia bowiem zadanie obliczenia całki na nieco tylko prostsze, w ogólnym przypadku, zagadnienie próbkowania z wielowymiarowego (zazwyczaj) rozkładu prawdopodobieństwa. Wyznaczenie całki wymaga wylosowania próby z rozkładu błędów oszacowań wartości zmiennych wejściowych $\boldsymbol{\delta}$.

Wydawać by się mogło, że można to uzyskać za pomocą dobrego generatora liczb losowych. Problem polega na tym, że generowanie liczb losowych to elementarne zadanie z zakresu metod numerycznych, ale tylko w przypadku jednowymiarowym, a w dodatku jedynie dla wybranych rozkładów prawdopodobieństwa. Natomiast w przypadku formuły całkowania (3.5.31) nie możemy próbkować oddzielnie wartości każdej z zaszumionych zmiennych wejściowych modelu prognostycznego, z ich jednowymiarowych rozkładów brzegowych. Błędy oszacowań wartości wejściowych na ogół nie będą bowiem niezależne. Przypomnijmy sobie chociażby wyznaczoną przez nas w poprzednim punkcie macierz kowariancji błędów temperatur maksymalnych i minimalnych, określoną przez (3.5.26). Niezbędne więc staje się wtedy znalezienie jakiejś efektywnej metody pozwalającej na próbkowanie z takiego rozkładu o wysokiej wymiarowości.

Pamiętajmy jednak, że w przypadku rozważanych przez nas problemów prognozy godzinnego zapotrzebowania na energię elektryczną (3.5.1) oraz prognozy maksymalnego godzinnego zapotrzebowania na energię (3.5.2) błędy prognoz temperatur $\boldsymbol{\delta}$ mają rozkład normalny $N(0, \mathbf{C}_{\boldsymbol{\delta}})$, o wartości oczekiwanej zero i macierzy kowariancji $\mathbf{C}_{\boldsymbol{\delta}}$ określonej przez zależność (3.5.26). Na szczęście wielowymiarowy rozkład normalny należy do wyjątków, ponieważ losowanie z niego próbek jest akurat operacją stosunkowo prostą (Brandt 1998).

Przyjmijmy ogólnie, że nasze zadanie polega na wylosowaniu próbki z k-wymiarowego rozkładu normalnego $N(\mathbf{a}, \mathbf{C}_{\boldsymbol{\delta}})$ zmiennych $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)$ o wartości oczekiwanej $\mathbf{a}$ i macierzy kowariancji $\mathbf{C}_{\boldsymbol{\delta}}$. Gęstość prawdopodobieństwa dla tego rozkładu dana jest wtedy za pomocą zależności:

$$p(\boldsymbol{\delta}) = c \exp\left\{-\frac{1}{2}(\boldsymbol{\delta}-\mathbf{a})^{\mathrm{T}} C_{\boldsymbol{\delta}}^{-1}(\boldsymbol{\delta}-\mathbf{a})\right\} \tag{3.5.32}$$

Macierz kowariancji $\mathbf{C}_{\boldsymbol{\delta}}$ jest z definicji symetryczna i dodatnio określona, podobnie macierz do niej odwrotna. Możemy wobec tego znaleźć jej rozkład Cholesky'ego na iloczyn macierzy trójkątnych, $\mathbf{C}_{\boldsymbol{\delta}}^{-1} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$. Jeżeli teraz zdefiniujemy nową zmienną $\mathbf{u} = \mathbf{L}(\boldsymbol{\delta} - \mathbf{a})$, to gęstość jej rozkładu możemy na podstawie (3.5.32) zapisać jako:

$$p(\mathbf{u}) = c\exp\left\{-\frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{u}\right\} = c\exp\left\{-\frac{1}{2}(u_1^2 + \ldots + u_k^2)\right\} = c\prod_{i=1}^{k}\exp\left\{-\frac{1}{2}u_i^2\right\} \qquad (3.5.33)$$

Jeśli spojrzymy na wzór (3.5.33), to widzimy, że gęstość prawdopodobieństwa w rozkładzie łącznym zmiennych u_i, $i = 1, \ldots, k$ równa jest iloczynowi gęstości prawdopodobieństw poszczególnych zmiennych. Oznacza to więc, że zmienne te są niezależne oraz mają standardowe rozkłady normalne, o wartości oczekiwanej 0 i odchyleniu standardowym 1.

Z przedstawionych rozważań wynika następujący prosty algorytm generowania próbek losowych z k-wymiarowego rozkładu normalnego $N(\mathbf{a}, \mathbf{C}_{\boldsymbol{\delta}})$.

Znajdujemy macierz odwrotną do macierzy kowariancji rozkładu $\mathbf{C}_{\boldsymbol{\delta}}$, a następnie macierz trójkątną dolną $\mathbf{L}$, z rozkładu Cholesky'ego macierzy $\mathbf{C}_{\boldsymbol{\delta}}^{-1}$. Są to standardowe operacje dostępne w każdym podstawowym pakiecie metod numerycznych lub łatwe do zaimplementowania samemu (Press, Teukolsky, Vetterling, Flannery 1992). Każdą kolejną k-wymiarową próbkę wylosowaną z rozkładu prawdopodobieństwa $N(\mathbf{a}, \mathbf{C}_{\boldsymbol{\delta}})$ tworzymy, powtarzając za każdym razem następujące operacje:

1. Generujemy k liczb u_i, $i = 1, \ldots, k$, o standardowym rozkładzie normalnym $N(0, 1)$. Możemy każdą z nich losować z rozkładu jednowymiarowego, ponieważ mają one być niezależne. To również jest dosyć standardowa operacja, zazwyczaj wykorzystuje się prostą metodę transformacji liczb losowych o rozkładzie jednostajnym (Press, Teukolsky, Vetterling, Flannery 1992; Brandt 1998). Sugeruje się przy tym raczej zastąpienie standardowych generatorów liczb losowych, dostępnych w środowiskach do budowy aplikacji, lepszymi rozwiązaniami (szczegóły – patrz: Press, Teukolsky, Vetterling, Flannery 1992; Brandt 1998).

2. Dla próbki $\mathbf{u} = (u_1, \ldots, u_k)$, złożonej z wylosowanych w poprzednim kroku liczb o rozkładzie $N(0, 1)$, tworzymy próbkę $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)$, o rozkładzie $N(\mathbf{a}, \mathbf{C}_{\boldsymbol{\delta}})$, korzystając ze wzoru:

$$\boldsymbol{\delta} = \mathbf{L}^{-1}\mathbf{u} + \mathbf{a} \qquad (3.5.34)$$

Pamiętajmy, że zaprezentowana metoda może być stosowana wyłącznie w przypadku próbkowania wielowymiarowego rozkładu normalnego. Jeżeli rzeczywisty rozkład błędu oszacowań wartości zaszumionych zmiennych

wejściowych modelu prognostycznego odchyla się w poważniejszy sposób od powziętego tu założenia albo wręcz ma postać niegaussowską, to metoda staje się bezużyteczna. Tak jak już wspomnieliśmy w przypadku ogólnym, zadanie próbkowania wielowymiarowego rozkładu nabiera znacznie bardziej złożonego charakteru.

Najczęściej stosowanym rozwiązaniem problemu pobierania próby z wielowymiarowego rozkładu prawdopodobieństwa są podejścia oparte na próbkowaniu Monte Carlo z wykorzystaniem łańcuchów Markowa. Do najważniejszych metod możemy tutaj zaliczyć Metropolis–Hastings oraz próbkowanie Gibbsa, które w zasadzie stanowi szczególny przypadek pierwszej metody. Obecnie przedstawimy jedynie zarys obu algorytmów. Szczegółowy przegląd zaawansowanych metod próbkowania rozkładów prawdopodobieństwa znaleźć można u MacKaya (2003) albo u Neala (1996).

Algorytm Metropolis–Hastings zastępuje bezpośrednie próbkowanie ze złożonego wielowymiarowego rozkładu prawdopodobieństwa o gęstości $p(\boldsymbol{\delta})$ generowaniem kolejnych próbek $\boldsymbol{\delta}^{t+1}$ w formie łańcucha Markowa, w którym wykorzystuje się próbki $\boldsymbol{\delta}'$ wylosowane z pewnego ustalonego rozkładu prawdopodobieństwa o gęstości $q(\boldsymbol{\delta}'; \boldsymbol{\delta}^t)$, zależnego od bieżącej próbki (stanu łańcucha) $\boldsymbol{\delta}^t$. Rozkład $q(\boldsymbol{\delta}'; \boldsymbol{\delta}^t)$ może być dowolnym ustalonym rozkładem prawdopodobieństwa, z którego potrafimy wylosować próbkę (MacKay 2003). Na przykład może to być rozkład normalny o wartości oczekiwanej w punkcie $\boldsymbol{\delta}^t$ i stałej kowariancji $\sigma^2\mathbf{I}$, gdzie przez $\mathbf{I}$ rozumiemy macierz tożsamościową.

Algorytm próbkowania Metropolis–Hastings możemy więc przedstawić w następującej formie. Rozpoczynamy od pewnego stanu początkowego łańcucha Markowa $\boldsymbol{\delta}^0$, zazwyczaj dobieranego w sposób losowy. Kolejne próbki dla $t = 0, 1, 2 \ldots, T$ generujemy, powtarzając kolejne kroki:

1. Mając stan łańcucha Markowa dla kroku t, następny stan, $t + 1$, wyznaczamy następująco: z proponowanego rozkładu $q(\boldsymbol{\delta}'; \boldsymbol{\delta}^t)$ losujemy próbkę $\boldsymbol{\delta}'$; aby ustalić następny stan łańcucha Markowa (kolejną próbkę) i wyznaczamy wielkość:

$$a = \frac{p(\boldsymbol{\delta}')}{p(\boldsymbol{\delta}^t)} \cdot \frac{q(\boldsymbol{\delta}^t; \boldsymbol{\delta}')}{q(\boldsymbol{\delta}'; \boldsymbol{\delta}^t)} \tag{3.5.35}$$

2. Nowy stan $\boldsymbol{\delta}^{t+1}$ ustalamy zgodnie z regułą:
jeżeli $a \geq 1$, to $\boldsymbol{\delta}^{t+1} = \boldsymbol{\delta}'$,
w przeciwnym razie:

$$\boldsymbol{\delta}^{t+1} = \begin{cases} \boldsymbol{\delta}' & \text{z prawdopodob. } a \\ \boldsymbol{\delta}^t & \text{z prawdopodob. } (1-a) \end{cases} \tag{3.5.36}$$

Zauważmy, że jeżeli proponowany rozkład prawdopodobieństwa jest symetryczny (jak np. w przypadku rozkładów gaussowskich), to drugi czynnik we wzorze (3.5.35) jest równy 1. Taką uproszczoną wersję tego algorytmu próbkowania nazywa się zazwyczaj krótko algorytmem Metropolis.

Algorytm próbkowania Gibbsa opiera się na założeniu, że jeżeli dany wielowymiarowy rozkład prawdopodobieństwa $p(\boldsymbol{\delta})$ jest zbyt złożony, by losować z niego próbki bezpośrednio, to możemy do generowania kolejnych próbek zastosować łańcuch Markowa stanów tworzonych poprzez losowanie po wartości kolejnych zmiennych próbki z ich rozkładów warunkowych $p(\delta_i / \delta_1, \ldots, \delta_{i-1}, \delta_{i+1}, \ldots, \delta_k)$, $i = 1, \ldots, k$ (MacKay 2003). Metodę tę można więc uznać za szczególny przypadek algorytmu Metropolis–Hastings, w którym wykorzystuje się rozkłady warunkowe poszczególnych zmiennych δ_i jako zastępcze rozkłady do faktycznego losowania próbek. Są one jednowymiarowe, możemy więc zazwyczaj poradzić sobie z nimi różnego rodzaju prostszymi metodami próbkowania. Oczywiście niezbędne jest wówczas wyznaczenie zależności warunkowych między zmiennymi modelu. Próbkowanie Gibbsa uznaje się więc za metodę prostszą od Metropolis–Hastings, ale o mniejszym zakresie zastosowania.

Podstawowe kroki algorytmu próbkowania Gibbsa możemy przedstawić w następujący sposób.

Podobnie jak w metodzie Metropolis–Hastings, również i w tym przypadku rozpoczynamy tworzenie łańcucha Markowa od pewnego stanu początkowego $\boldsymbol{\delta}^0$, zazwyczaj dobieranego w sposób losowy. Kolejne próbki dla $t = 0, 1, 2 \ldots, T$ generujemy w następujący sposób:

Na podstawie stanu łańcucha Markowa t stan $t + 1$ otrzymujemy następująco: generujemy po jednej zmiennej na raz z rozkładu warunkowego tej zmiennej względem pozostałych, aktualizując wartości tych zmiennych natychmiast po ich uzyskaniu. Nową próbkę $\boldsymbol{\delta}^{t+1}$ otrzymujemy więc, losując po kolei:

$$\begin{aligned}
&\delta_1^{t+1} \sim p(\delta_1 / \delta_2^t, \delta_3^t, \ldots, \delta_k^t) \\
&\delta_2^{t+1} \sim p(\delta_2 / \delta_1^{t+1}, \delta_3^t, \ldots, \delta_k^t) \\
&\ldots \\
&\delta_i^{t+1} \sim p(\delta_i / \delta_1^{t+1}, \ldots, \delta_{i-1}^{t+1}, \delta_{i+1}^t, \ldots, \delta_k^t) \\
&\ldots \\
&\delta_k^{t+1} \sim p(\delta_k / \delta_1^{t+1}, \ldots, \delta_{k-1}^{t+1})
\end{aligned} \tag{3.5.37}$$

Można pokazać, że zarówno w przypadku metody Metropolis–Hastings, jak i próbkowania Gibbsa, rozkład prawdopodobieństwa, z którego próbkujemy $\boldsymbol{\delta}^t$, dąży do $p(\boldsymbol{\delta})$, przy $t \to \infty$. Problem jednak polega na tym, że kolejne próbki $\boldsymbol{\delta}^t$ tworzymy za pomocą łańcucha Markowa, nie są więc niezależne. Pozostają powiązane między sobą z niewielkim, ale jednak większym od zera, prawdopo-

dobieństwem. W związku z tym, wykorzystując je, pamiętać musimy o dwóch następujących kwestiach:

– dla obydwu metod pewna liczba początkowo wygenerowanych próbek musi zostać zignorowana, ponieważ zależne są one od wartości początkowej; okres ten, określany często jako „wypalanie" (*burn-in*), wykorzystany może zostać również do oszacowania różnego rodzaju parametrów (np. parametru σ, jeżeli w metodzie Metropolis zastosujemy rozkład Gaussa); typowo przyjmuje się, że odrzucamy około 1000 próbek, ale zasadniczo liczba ta zależy od konkretnego przypadku,

– ponieważ kolejne próbki wykazują pewną korelację, w związku z tym do finalnej puli nie wybiera się wszystkich kolejnych wygenerowanych wzorców; np. ostatecznie akceptuje się co setny wygenerowany stan łańcucha Markowa, pozostałe pomijając.

3.5.4. Aproksymacja gęstości prawdopodobieństwa niepewności wejść modelu

Rozważane w poprzednim punkcie metody wyznaczania wartości oczekiwanej predykcji zapotrzebowania na energię $E(y / \mathbf{z}, \mathbf{x}^d)$ w warunkach niepewności wejść określonej przy użyciu zależności (3.5.28) nie zawsze mogą zostać zastosowane. Metoda wyznaczania całki opartej na próbkowaniu Monte Carlo (3.5.31) ma przynajmniej dwa poważne ograniczenia. Przede wszystkim niezbędna jest naturalnie znajomość rozkładu prawdopodobieństwa $p(\mathbf{x}^z / \mathbf{z})$, określającego możliwe wartości nieznanych zmiennych $\mathbf{x}^z$ przy danych ich oszacowaniach $\mathbf{z}$ (lub rozkładu błędu tego oszacowania $p(\boldsymbol{\delta})$). Ponadto musimy tutaj wspomnieć o kwestii efektywności tego rozwiązania. Problem niskiej efektywności stanowi zasadniczą bolączkę wszelkich metod Monte Carlo. Również i w naszym przypadku, nawet jeżeli znamy rozkład $p(\mathbf{x}^z / \mathbf{z})$, to jego próbkowanie może być czynnością dosyć kosztowną obliczeniowo.

Uzyskanie dostatecznie wysokiej dokładności wartości całki za pomocą zależności (3.5.31) wymagać może wylosowania próby o znacznej liczebności. Problem ten potęguje się zwłaszcza przy dużej liczbie szacowanych zmiennych wejściowych, wykorzystanej do prognozy sieci neuronowej (neuronowo-rozmytej). Zwiększanie rozmiarów próbkowanej przestrzeni wejściowej powoduje konieczność tworzenia coraz większych prób, które by ją właściwie reprezentowały. Nawet jeśli rozkład błędu $p(\boldsymbol{\delta})$ ma charakter gaussowski, może to spowodować powstanie nieakceptowanych czasów szacowania wartości oczekiwanej $E(y / \mathbf{z}, \mathbf{x}^d)$. W przypadku konieczności zastosowania metod próbkowania Monte Carlo, w których wykorzystuje się łańcuchy Markowa – nienależące same z siebie do najbardziej efektywnych – o podobne problemy jest szczególnie łatwo.

Rozwiązanie omijające obydwa wskazane problemy polega na zastosowaniu metody obliczania wartości oczekiwanej prognozy $E(y / \mathbf{z}, \mathbf{x}^d)$, zastępującej we wzorze (3.5.28) gęstość rozkładu prawdopodobieństwa $p(\mathbf{x}^z / \mathbf{z})$ jej odpowiednią aproksymantą, dla której można w łatwiejszy sposób wyznaczyć wartość całki. Nie musimy wtedy znać dokładnej postaci $p(\mathbf{x}^z / \mathbf{z})$ i, niejako przy okazji, rozwiązuje się problem całkowania. Warunek stanowi, rzecz jasna, znalezienie dokładnej i szybkiej metody aproksymacji rozważanego rozkładu prawdopodobieństwa.

Podejściem, które możemy tutaj zastosować, jest tzw. aproksymacja Parzena (nazywana również aproksymacją oknami Parzena). Polega ona na przybliżaniu wartości funkcji gęstości prawdopodobieństwa rozkładu pewnego zbioru danych za pomocą gaussowskich okien scentrowanych wokół tychże punktów.

Jeżeli mamy zbiór danych wartości pewnej zmiennej losowej $\mathbf{x}_k$, $k = 1, \ldots, N$, to funkcję gęstości rozkładu prawdopodobieństwa tej zmiennej możemy aproksymować przy użyciu zależności:

$$p(\mathbf{x}) \approx \frac{1}{N} \sum_{k=1}^{N} G(\mathbf{x}, \mathbf{x}_k, \sigma) \tag{3.5.38}$$

gdzie $G(\mathbf{x}, \mathbf{x}_k, \sigma)$ jest wielowymiarową krzywą Gaussa o środku w punkcie $\mathbf{x}_k$ i odchyleniu standardowym σ:

$$G(\mathbf{x}, \mathbf{x}_k, \sigma) = m \exp\left(-\frac{1}{2\sigma^2} \| \mathbf{x} - \mathbf{x}_k \|^2 \right) \tag{3.5.39}$$

Współczynnik m ma za zadanie zapewniać właściwą normalizację krzywej Gaussa do jednostkowej całki, tak by spełniała warunki dla funkcji gęstości prawdopodobieństwa. W naszym jednak przypadku jest on nieistotny, ponieważ, jak zobaczymy dalej, zostanie zredukowany. Będziemy więc przyjmować jego wartość jako równą 1.

Idea aproksymacji za pomocą okien Parzena jest dosyć oczywista. Dla nowego wzorca $\mathbf{x}$, jeżeli w niewielkiej odległości od niego położonych jest wiele punktów danych, wartość funkcji gęstości prawdopodobieństwa będzie duża. Jeżeli punkt danych znajduje się w regionie przestrzeni wartości zmiennej losowej o mniejszej liczbie wzorców dalej położonych, to gęstość prawdopodobieństwa będzie dużo niższa.

Zbiorem danych, który wykorzystywać będziemy do aproksymacji Parzena, będzie zbiór treningowy dla sieci neuronowej (neuronowo-rozmytej) $D = \{\mathbf{x}_k, y_k\}$, $k = 1, \ldots, N$. Pamiętajmy, że wzorce wejściowe każdej obserwacji występującej w zbiorze D podzielić możemy na dwie grupy $\mathbf{x}_k = (\mathbf{x}_k^d, \mathbf{x}_k^z)$, gdzie $\mathbf{x}_k^d$ są wartościami zn zmiennych wejściowych znanych dokładnie, natomiast

$\mathbf{x}_k^z$ wartościami *sz* zmiennych zaszumionych, które w danych treningowych znane są, co prawda, dokładnie, ale podczas prognozowania znane będą jedynie w przybliżeniu, za pomocą oszacowań $\mathbf{z}$. Oczywiście $zn + sz = n$, gdzie n jest liczbą wszystkich zmiennych wejściowych.

Korzystając z zależności (3.5.28) (lub wcześniejszej (3.5.11)) oraz z definicji prawdopodobieństwa warunkowego, możemy wartość oczekiwaną prognozy dla zaszumionych danych wejściowych zapisać w następujący sposób:

$$E(y/\mathbf{z},\mathbf{x}^d) = \int_{-\infty}^{\infty} f(\mathbf{x}^z,\mathbf{x}^d,\mathbf{w})p(\mathbf{x}^z/\mathbf{z})d\mathbf{x}^z = \\ = \frac{1}{p(\mathbf{z})}\int_{-\infty}^{\infty} f(\mathbf{x}^z,\mathbf{x}^d,\mathbf{w})p(\mathbf{x}^z,\mathbf{z})d\mathbf{x}^z \tag{3.5.40}$$

gdzie $p(\mathbf{x}^z, \mathbf{z})$ jest funkcją gęstości rozkładu wspólnego prawdopodobieństwa możliwych wartości zmiennych $\mathbf{x}^z$ oraz ich znanych oszacowań $\mathbf{z}$, zaś $p(\mathbf{z})$ rozkładem brzegowym oszacowań $\mathbf{z}$. Gęstości obydwu rozkładów prawdopodobieństwa będziemy aproksymować za pomocą okien Parzena scentrowanych wokół wzorców danych $\mathbf{x}_k^z$ pochodzących ze zbioru treningowego sieci:

$$p(\mathbf{z}) \approx \frac{1}{N}\sum_{k=1}^{N} G(\mathbf{z},\mathbf{x}_k^z,\sigma) \tag{3.5.41a}$$

$$p(\mathbf{x}^z,\mathbf{z}) \approx \frac{1}{N}\sum_{k=1}^{N} G((\mathbf{x}^z,\mathbf{z}),(\mathbf{x}_k^z,\mathbf{x}_k^z),\sigma) \tag{3.5.41b}$$

Zwróćmy uwagę, że w przypadku rozkładu brzegowego (3.5.41a) do aproksymacji używamy *sz* wymiarowych funkcji Gaussa postaci (3.5.39), zaś w przypadku rozkładu wspólnego (3.5.41b) – funkcji o wymiarze 2*sz*. Oczywiście *sz* jest, zgodnie z wcześniejszym oznaczeniem, liczbą szacowanych zmiennych wejściowych. Możemy wówczas zapisać (3.5.40) w następujący sposób:

$$E(y/\mathbf{z},\mathbf{x}^d) \approx \frac{1}{\sum_{k=1}^{N} G(\mathbf{z},\mathbf{x}_k^z,\sigma)}\int_{-\infty}^{\infty} f(\mathbf{x}^z,\mathbf{x}^d,\mathbf{w})\sum_{k=1}^{N} G((\mathbf{x}^z,\mathbf{z}),(\mathbf{x}_k^z,\mathbf{x}_k^z),\sigma)d\mathbf{x}^z = \tag{3.5.42}$$

$$= \frac{1}{\sum_{k=1}^{N} G(\mathbf{z},\mathbf{x}_k^z,\sigma)}\sum_{k=1}^{N}\int_{-\infty}^{\infty} f(\mathbf{x}^z,\mathbf{x}^d,\mathbf{w})G((\mathbf{x}^z,\mathbf{z}),(\mathbf{x}_k^z,\mathbf{x}_k^z),\sigma)d\mathbf{x}^z$$

Zauważmy, że w wyrażeniach podcałkowych w zależności (3.5.42) funkcja realizowana przez sieć neuronową mnożona jest przez wartości funkcji okna gaussowskiego. W obszarach przestrzeni całkowania, w których dana funkcja okna ma wartości niemal równe zero, odpowiednie wyrażenie podcałkowe również więc niemal zanika. Dla wyznaczenia wartości każdej całki wartość funkcji sieci neuronowej (neuronowo-rozmytej) $f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w})$ względem zaszumionych zmiennych $\mathbf{x}^z$ okazuje się zatem istotna jedynie w obszarze „szerokości" okna, w którym funkcja okna jest istotnie różna od 0. Jeżeli więc przyjmiemy aproksymację funkcji f za pomocą funkcji obszarami stałej (wielowymiarowego odpowiednika funkcji „schodkowej"):

$$f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) \approx f(\mathbf{x}^z_k, \mathbf{x}^d, \mathbf{w}) \text{ w obszarze każdego okna } G((\mathbf{x}^z, \mathbf{z}), (\mathbf{x}^z_k, \mathbf{x}^z_k), \sigma) \quad (3.5.43)$$

czyli zastępując wyjście sieci w obrębie każdego z okien stałą wartością tej sieci w środku tego okna, błąd w obliczonej całce nie powinien być duży, zwłaszcza przy znacznej liczbie wzorców danych w zbiorze treningowym, co pozwala na redukcję szerokości okien σ niezbędnych do pokrycia przestrzeni danych. Dla każdej całki występującej w (3.5.42) możemy więc napisać:

$$\begin{aligned}
&\int_{-\infty}^{\infty} f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) G((\mathbf{x}^z, \mathbf{z}), (\mathbf{x}^z_k, \mathbf{x}^z_k), \sigma) d\mathbf{x}^z \approx \\
&\qquad \approx \int_{-\infty}^{\infty} f(\mathbf{x}^z_k, \mathbf{x}^d, \mathbf{w}) G((\mathbf{x}^z, \mathbf{z}), (\mathbf{x}^z_k, \mathbf{x}^z_k), \sigma) d\mathbf{x}^z = \\
&\qquad = f(\mathbf{x}^z_k, \mathbf{x}^d, \mathbf{w}) \int_{-\infty}^{\infty} G((\mathbf{x}^z, \mathbf{z}), (\mathbf{x}^z_k, \mathbf{x}^z_k), \sigma) d\mathbf{x}^z
\end{aligned} \quad (3.5.44)$$

Zauważmy teraz, że całkując okno gaussowskie $G((\mathbf{x}^z, \mathbf{z}), (\mathbf{x}^z_k, \mathbf{x}^z_k), \sigma)$ o wymiarze $2sz$ względem nieznanych zmiennych wejściowych modelu prognostycznego $\mathbf{x}^z$, rzutujemy je na podprzestrzeń definiowaną przez znane oszacowania tych zmiennych $\mathbf{z}$ i otrzymujemy w wyniku tej operacji krzywą Gaussa o zredukowanym wymiarze sz:

$$\int_{-\infty}^{\infty} G((\mathbf{x}^z, \mathbf{z}), (\mathbf{x}^z_k, \mathbf{x}^z_k), \sigma) d\mathbf{x}^z = G(\mathbf{z}, \mathbf{x}^z_k, \sigma) \quad (3.5.45)$$

czyli podstawiając (3.5.45) do wzoru (3.5.44), otrzymujemy:

$$\int_{-\infty}^{\infty} f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) G((\mathbf{x}^z, \mathbf{z}), (\mathbf{x}^z_k, \mathbf{x}^z_k), \sigma) d\mathbf{x}^z \approx f(\mathbf{x}^z_k, \mathbf{x}^d, \mathbf{w}) G(\mathbf{z}, \mathbf{x}^z_k, \sigma) \quad (3.5.46)$$

Ostatecznie więc wykorzystując zależność (3.5.46) w zależności (3.5.42), otrzymujemy finalny wzór na aproksymację Parzena wartości oczekiwanej prognozy $E(y / \mathbf{z}, \mathbf{x}^d)$ dla znanych oszacowań zmiennych wejściowych $\mathbf{z}$:

$$E(y / \mathbf{z}, \mathbf{x}^d) \approx \frac{\sum_{k=1}^{N} f(\mathbf{x}_k^z, \mathbf{x}^d, \mathbf{w}) G(\mathbf{z}, \mathbf{x}_k^z, \sigma)}{\sum_{k=1}^{N} G(\mathbf{z}, \mathbf{x}_k^z, \sigma)} \tag{3.5.47}$$

Kilka słów musimy poświęcić kwestii doboru parametru σ. Jego wartość nie może być zbyt mała, ponieważ nie będzie wtedy zapewniał odpowiedniego pokrycia aproksymowanej przestrzeni oknami Parzena. Nie może być również za duża, aby nie zniekształcać wyników aproksymacji gęstości prawdopodobieństwa wpływem daleko położonych wzorców danych. W naszym przypadku dobraliśmy wartość parametru σ na podstawie średniej odległości między wzorcami danych w zbiorze treningowym sieci neuronowej (neuronowo-rozmytej) $\mathbf{x}^z_k$., $k = 1, \ldots, N$. Tresp, Neuneier, Ahmad (1995) proponują podejście polegające na odrzuceniu w każdym kroku i-tego wzorca danych i oszacowaniu związanej z nim gęstości prawdopodobieństwa za pomocą aproksymacji oknami Parzena dla pozostałych wzorców danych:

$$p(\mathbf{x}_i^z) \approx \frac{1}{N-1} \sum_{k=1, k\neq i}^{N} G(\mathbf{x}_i^z, \mathbf{x}_k^z, \sigma) \tag{3.5.48}$$

Następnie wartość σ określamy, znajdując maksimum zlogarytmowanej funkcji wiarygodności $\Sigma_i\, p(\mathbf{x}_i^z)$. Zauważmy jeszcze, że przyjmując stałą wartość σ, tak jak już wcześniej wspominaliśmy, możemy we wzorze funkcji Gaussa (3.5.39) przyjąć współczynnik normalizacyjny $m = 1$, ponieważ podczas aproksymacji za pomocą formuły (3.5.47) i tak współczynniki te się redukują.

Oszacowanie wartości oczekiwanej prognozy $E(y / \mathbf{z}, \mathbf{x}^d)$ dla znanych oszacowań zmiennych wejściowych modelu $\mathbf{z}$ za pomocą formuły (3.5.47) pozwala nam uniknąć wspomnianych wcześniej problemów z wykorzystywanymi w poprzednim punkcie metodami całkowania, w których stosuje się próbkowanie Monte Carlo. Nie musimy określać rozkładu prawdopodobieństwa definiującego niepewność szacowanych wartości zmiennych wejściowych sieci neuronowej (neuronowo-rozmytej) $\mathbf{z}$, aproksymując ten rozkład za pomocą okien Parzena. Co więcej, nie musimy przeprowadzać, często bardzo pracochłonnego, próbkowania rozkładu możliwych wartości tych zmiennych. Jako odpowiednia próba służą nam tutaj po prostu dane treningowe. Modelujemy zachowanie prognozy dla różnych wartości niepewnych zmiennych wejściowych przy użyciu analizy wyjścia modelu dla (znanych dokładnie) wartości tych zmiennych występują-

cych we wzorcach treningowych $\mathbf{x}^z_k$. W formule (3.5.47) ważmy tylko udział każdej obserwacji treningowej w końcowym wyniku, w zależności od znormalizowanej odległości między tą obserwacją a oszacowaniami wartości tych zmiennych **z**.

Podobnie jak w poprzednim rozdziale, dla metod Monte Carlo przetestujemy obecnie działanie formuły wyznaczania wartości oczekiwanej prognozy $E(y / \mathbf{z}, \mathbf{x}^d)$ (3.5.31), weryfikując praktycznie jej działanie dla przedstawionych w punkcie 3.5.1 zagadnień krótkoterminowego prognozowania zapotrzebowania na energię elektryczną. Przypomnijmy, że jako zadania testowe określiliśmy tam dwa problemy: pierwszy dotyczący prognozy godzinnego zapotrzebowania na energię elektryczną, z dwudniowym wyprzedzeniem czasowym (zależność (3.5.1)), oraz drugi dotyczący prognozy maksymalnego godzinnego zapotrzebowania na energię, z dwudniowym wyprzedzeniem czasowym (zależność (3.5.2)) (Bartkiewicz 2000c, e). Badaną techniką prognozowania wykorzystywaną w obydwu przypadkach są warstwowe sieci perceptronowe MLP. Otrzymane wyniki prognozy przedstawione zostały odpowiednio w tabeli 3.5.7 oraz w tabeli 3.5.8.

Tabela 3.5.7. Porównanie dokładności prognoz godzinnego zapotrzebowania na energię dla dokładnych wartości temperatur i metodą aproksymacji Parzena

Godz.	Temperatury dokładne		Aproksymacja Parzena		Godz.	Temperatury dokładne		Aproksymacja Parzena	
	MAE (kWh)	MAPE (%)	MAE (kWh)	MAPE (%)		MAE (kWh)	MAPE (%)	MAE (kWh)	MAPE (%)
1	6 262	2,99	6 062	2,91	13	6 429	2,57	7 711	3,08
2	4 863	2,39	4 999	2,47	14	5 600	2,25	7 486	2,97
3	4 438	2,24	4 622	2,34	15	6 450	2,52	8 288	3,21
4	3 970	2,01	4 350	2,21	16	9 196	3,60	10 406	4,07
5	4 130	2,06	4 479	2,24	17	8 385	3,19	9 344	3,56
6	4 544	2,21	4 884	2,38	18	7 918	2,89	8 374	3,02
7	6 292	2,71	6 580	2,85	19	7 482	2,69	7 488	2,66
8	6 642	2,71	6 984	2,85	20	4 909	1,61	5 307	1,73
9	6 239	2,51	6 594	2,65	21	5 822	1,99	6 320	2,16
10	6 631	2,61	7 031	2,76	22	6 203	2,36	6 273	2,39
11	7 153	2,81	7 764	3,06	23	5 078	2,07	5 323	2,17
12	6 093	2,40	7 211	2,84	24	5 481	2,41	5 494	2,43
Średnia ze wszystkich godzin						6 092	2,49	6 641	2,71

Źródło: opracowanie własne.

Tabela 3.5.8. Porównanie dokładności prognoz maksymalnego godzinnego zapotrzebowania na energię dla dokładnych wartości temperatur i metodą aproksymacji Parzena

Metoda	MAE (kWh)	MAX AE (kWh)	MAPE (%)	MAX APE (%)
Temperatury dokładne	7 918	29 550	2,88	16,75
Aproksymacja Parzena	8 336	30 005	3,00	16,99

Źródło: opracowanie własne.

Analizując dane znajdujące się w tabelach 3.5.7 oraz 3.5.8, widzimy, że w badanych przypadkach wyniki zastosowania aproksymacji Parzena do wyznaczania wartości oczekiwanej prognozy $E(y / \mathbf{z}, \mathbf{x}^d)$ są bardzo podobne jak w przypadku metod opartych na całkowaniu z wykorzystaniem próbkowania Monte Carlo. Jeśli porównujemy błędy prognoz z danymi zaprezentowanymi odpowiednio w tabelach 3.5.5 i 3.5.6, widzimy, że są one niemal identyczne. Pozwala to wysnuć wniosek, że dla rozważanych zadań obie metody mogą być stosowane wymiennie.

Oczywiście oznacza to również, że w testowanych przypadkach obie metody nie pozwoliły na znaczące odzyskanie dokładności wynikającej z błędów prognoz temperatury minimalnej i maksymalnej podawanych na wejściu modelu. Różnica w stosunku do bezpośredniego użycia oszacowań tych zmiennych (tabela 3.5.1 i 3.5.2) jest niewielka, rzędu kilku setnych procent. Efekt taki nie jest jednak niczym niespodziewanym. Tak jak przewidywaliśmy, lokalna linearyzacja modelu neuronowego lub neuronowo-rozmytego dla stosunkowo niewielkich błędów prognoz temperatur w krótkim horyzoncie czasowym nie powoduje znaczniejszego obciążenia prognozy. Ponownie jednak zwróćmy uwagę na fakt, że modele zapotrzebowania na energię, w których wykorzystuje się w szerszym zakresie zmienne wejściowe definiujące stan warunków atmosferycznych (takie jak ciśnienie atmosferyczne, wilgotność powietrza lub siła wiatru), powinny wykazywać większą wrażliwość na błędy ich prognoz.

3.5.5. Uproszczone rozwiązania dla przypadków szczególnych

W metodach wyznaczania optymalnej wartości prognozy (wartości oczekiwanej) w warunkach niepewności części zmiennych wejściowych, prezentowanych w punktach 3.4.3 i 3.4.4, wykorzystuje się mechanizm próbkowania przestrzeni błędu wartości tych zmiennych za pomocą metod Monte Carlo czy też przy użyciu danych ze zbioru treningowego. Oba te podejścia nieodmiennie charakteryzują się stosunkowo niską efektywnością obliczeniową i niewielką kontrolą nad dokładnością otrzymywanych wyników. Znalezienie wartości oczekiwanej prognozy wymaga bowiem scałkowania wyjścia modelu predykcji

względem zazwyczaj wielowymiarowego rozkładu prawdopodobieństwa opisującego niepewność danych wejściowych.

Problem, z którym stykamy się podczas próby rozwiązania tej kwestii, stanowi, jak już zresztą wcześniej wspominaliśmy, praktyczny niedobór dobrych metod numerycznych, które pozwoliłyby na wyznaczanie złożonych wielowymiarowych całek, podobnych do tej występującej w zależności (3.5.28) (lub alternatywnie (3.5.11)). Sytuacja tutaj jest zupełnie odmienna niż w przypadku jednowymiarowym, gdzie mamy do czynienia z prawdziwym bogactwem różnego rodzaju kwadratur o wysokiej efektywności działania algorytmu i szerokim spektrum możliwości zastosowań.

Dla niektórych przypadków szczególnych, związanych z określoną architekturą modelu neuronowego (neuronowo-rozmytego) oraz kształtu błędu oszacowań wartości wejściowych, istnieją pewne możliwości sprowadzenia zadania znajdowania wartości oczekiwanej prognozy względem wielowymiarowego rozkładu prawdopodobieństwa wejść do wyznaczenia kilku całek jednowymiarowych. Na zakończenie podrozdziału 3.5 przedstawimy, jako przykład tego typu rozwiązań, podejście zaproponowane przez Dinga (1999). Od razu jednak wprowadźmy istotne zastrzeżenie: zaprezentujemy tutaj jedynie zarys tej metody – wyłącznie w celach ilustracyjnych, tak by wskazać zainteresowanemu Czytelnikowi pewne możliwości ewentualnych dalszych dociekań w tej dziedzinie. Nie przeprowadzaliśmy również żadnej weryfikacji praktycznej proponowanego podejścia w zadaniach krótkoterminowej prognozy zapotrzebowania na energię (chociaż sam Ding pokazuje pewne wyniki także i w tej dziedzinie – patrz: Ding 1999).

Proponowana metoda dotyczy warstwowych sieci perceptronowych z jedną warstwą ukrytą i jednym neuronem wyjściowym o liniowej funkcji aktywacji. Zauważmy, że dostosowując ogólną formułę przetwarzania sieci perceptronowej, określoną w punkcie 2.2.2 za pomocą zależności (2.2.1), do obecnego przypadku, równanie przetwarzające sieci możemy zapisać:

$$f(\mathbf{x},\mathbf{w}) = \sum_{i=1}^{h} w_i^{(2)} \varphi\left((\mathbf{w}_i^{(1)})^{\mathrm{T}} \mathbf{x} \right) = \sum_{i=1}^{h} w_i^{(2)} \varphi\left(\sum_{j=1}^{n} w_{ij}^{(1)} x_j \right) \tag{3.5.49}$$

gdzie h jest liczbą neuronów w warstwie ukrytej, n liczbą wejść, $w_{ij}^{(1)}$ współczynnikiem wagowym j-tego wejścia, i-tego neuronu w warstwie ukrytej, $w_i^{(2)}$ jest współczynnikiem wagowym i-tego wejścia jedynego neuronu wyjściowego, natomiast φ funkcją aktywacji neuronów, zaś $\mathbf{w}$ zbiorem wszystkich parametrów (wag) sieci.

Przypomnijmy, że wejścia sieci podzielić możemy na dwie grupy: zn zmiennych wejściowych $\mathbf{x}^d$, znanych dokładnie, oraz sz zmiennych $\mathbf{x}^z$ zaszumionych (gdzie $zn + sz = n$), znanych w postaci oszacowań $\mathbf{z}$. Oznaczmy dla każdego i-tego neuronu w warstwie ukrytej oddzielnie: wagi odpowiadające znanym wejściom $\mathbf{x}^d$ przez $\boldsymbol{\omega}_i$, zaś wagi odpowiadające wejściom niepewnym $\mathbf{x}^z$

przez $\mathbf{v}_i$. Wówczas formułę działania sieci neuronowej (3.5.49) możemy przepisać jako:

$$f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) = \sum_{i=1}^{h} w_i^{(2)} \varphi\left(\boldsymbol{\omega}_i^{\mathrm{T}} \mathbf{x}^d + \mathbf{v}_i^{\mathrm{T}} \mathbf{x}^z\right) \tag{3.5.50}$$

Zakładamy również, że błąd $\boldsymbol{\delta}$ oszacowań nieznanych zmiennych wejściowych $\mathbf{x}^z$ jest addytywnym szumem losowym, tj. $\mathbf{x}^z = \mathbf{z} + \boldsymbol{\delta}$, przy czym ma on wielowymiarowy rozkład normalny $N(0, \mathbf{C}_{\boldsymbol{\delta}})$ o wartości oczekiwanej zero i pewnej macierzy kowariancji $\mathbf{C}_{\boldsymbol{\delta}}$.

Wróćmy teraz do naszej zależności na wartość oczekiwaną prognozy zapotrzebowania na energię $E(y / \mathbf{z}, \mathbf{x}^d)$, dla niepewnych wartości zmiennych wejściowych $\mathbf{z}$, określoną za pomocą wzoru (3.5.28) (lub alternatywnie (3.5.11)). Dla sieci perceptronowej o strukturze (3.5.50) możemy zapisać tę formułę w następujący sposób:

$$\begin{aligned} E(y / \mathbf{z}, \mathbf{x}^d) &= \int_{-\infty}^{\infty} f(\mathbf{x}^z, \mathbf{x}^d, \mathbf{w}) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z = \\ &= \int_{-\infty}^{\infty} \left(\sum_{i=1}^{h} w_i^{(2)} \varphi\left(\boldsymbol{\omega}_i^{\mathrm{T}} \mathbf{x}^d + \mathbf{v}_i^{\mathrm{T}} \mathbf{x}^z\right) \right) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z = \\ &= \sum_{i=1}^{h} w_i^{(2)} \int_{-\infty}^{\infty} \varphi\left(\boldsymbol{\omega}_i^{\mathrm{T}} \mathbf{x}^d + \mathbf{v}_i^{\mathrm{T}} \mathbf{x}^z\right) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z \end{aligned} \tag{3.5.51}$$

Rozkład prawdopodobieństwa nieznanych prawdziwych wartości zmiennych $\mathbf{x}^z$, przy znanych ich oszacowaniach $\mathbf{z}$, generowany jest przez rozkład błędu tego oszacowania $\boldsymbol{\delta}$. Podstawiając teraz do (3.5.51) $\mathbf{x}^z = \mathbf{z} + \boldsymbol{\delta}$, otrzymujemy więc:

$$\begin{aligned} E(y / \mathbf{z}, \mathbf{x}^d) &= \sum_{i=1}^{h} w_i^{(2)} \int_{-\infty}^{\infty} \varphi\left(\boldsymbol{\omega}_i^{\mathrm{T}} \mathbf{x}^d + \mathbf{v}_i^{\mathrm{T}} \mathbf{x}^z\right) p(\mathbf{x}^z / \mathbf{z}) d\mathbf{x}^z = \\ &= \sum_{i=1}^{h} w_i^{(2)} \int_{-\infty}^{\infty} \varphi\left(\boldsymbol{\omega}_i^{\mathrm{T}} \mathbf{x}^d + \mathbf{v}_i^{\mathrm{T}} \mathbf{z} + \mathbf{v}_i^{\mathrm{T}} \boldsymbol{\delta}\right) p(\boldsymbol{\delta}) d\boldsymbol{\delta} \end{aligned} \tag{3.5.52}$$

gdzie $p(\boldsymbol{\delta})$ jest gęstością rozkładu prawdopodobieństwa błędu $\boldsymbol{\delta}$, czyli, jak założyliśmy wcześniej, gęstością wielowymiarowego rozkładu normalnego $N(0, \mathbf{C}_{\boldsymbol{\delta}})$ o wartości oczekiwanej zero i macierzy kowariancji $\mathbf{C}_{\boldsymbol{\delta}}$.

Zauważmy teraz jedną rzecz. Otóż część pobudzenia neuronów warstwy ukrytej sieci związana z błędami zmiennych wejściowych, czyli wartość $\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\delta}$, jest skalarną zmienną losową, a ponadto stanowi ona wynik przekształcenia liniowego błędu $\boldsymbol{\delta}$. Jeżeli więc $\boldsymbol{\delta}$ ma wielowymiarowy rozkład normalny $N(0, \mathbf{C}_{\boldsymbol{\delta}})$, to $\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\delta}$ ma także rozkład normalny, lecz jednowymiarowy $N(0, \sqrt{\mathbf{v}_i^{\mathrm{T}}\mathbf{C}_{\boldsymbol{\delta}}\mathbf{v}_i})$. Wariancję rozkładu pobudzenia $\mathbf{v}_i^{\mathrm{T}}\mathbf{C}_{\boldsymbol{\delta}}\mathbf{v}_i$ neuronu otrzymujemy bezpośrednio, korzystając z prawa propagacji błędów (patrz zależność (3.2.28) w punkcie 3.2.4).

Wprowadźmy więc nową zmienną t, określającą standaryzowane (do jednostkowego odchylenia standardowego) pobudzenie neuronów warstwy ukrytej sieci, związane z błędami zmiennych wejściowych:

$$t = \frac{\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\delta}}{\sqrt{\mathbf{v}_i^{\mathrm{T}}\mathbf{C}_{\boldsymbol{\delta}}\mathbf{v}_i}} \tag{3.5.53}$$

Zmienna t ma, oczywiście, rozkład normalny standardowy $N(0, 1)$. Podstawiając ją teraz do (3.5.52), otrzymujemy ostateczną formułę pozwalającą na wyznaczenie wartości oczekiwanej prognozy $E(y / \mathbf{z}, \mathbf{x}^d)$, dla niepewnych oszacowań wartości zmiennych wejściowych $\mathbf{z}$, w przypadku warstwowych sieci perceptronowych o architekturze (3.5.49) i normalnych błędów $\boldsymbol{\delta}$:

$$E(y / \mathbf{z}, \mathbf{x}^d) = \sum_{i=1}^{h} w_i^{(2)} \int_{-\infty}^{\infty} \varphi\left(\boldsymbol{\omega}_i^{\mathrm{T}}\mathbf{x}^d + \mathbf{v}_i^{\mathrm{T}}\mathbf{z} + \sqrt{\mathbf{v}_i^{\mathrm{T}}\mathbf{C}_{\boldsymbol{\delta}}\mathbf{v}_i}\, t \right) p(t)dt \tag{3.5.54}$$

gdzie $p(t)$ oznacza już obecnie funkcję gęstości normalnego standardowego rozkładu prawdopodobieństwa, $N(0, 1)$.

Całki występujące we wzorze (3.5.54) są już zwykłymi całkami jednowymiarowymi. Można je bez trudu i efektywnie obliczyć, stosując standardowe kwadratury numeryczne, stosunkowo łatwe do samodzielnej implementacji bądź osiągalne w powszechnie dostępnych pakietach i narzędziach obliczeniowych.

3.6. Podsumowanie

W bieżącym rozdziale przyjrzeliśmy się metodom modelowania niepewności prognoz krótkoterminowego zapotrzebowania na energię elektryczną. Przeprowadzona została analiza i interpretacja wyjścia modelu prognostycznego oraz źródeł jego niepewności, które należy uwzględnić przy modelowaniu prognozy. Rozważane w pracy predyktory – w formie sieci neuronowych

i neuronowo-rozmytych – należą do modeli regresyjnych dopasowywanych do danych metodą najmniejszych kwadratów. Wyjście tego rodzaju modelu (prognoza) może być interpretowane jako wartość oczekiwana zapotrzebowania na energię, dla danego wzorca danych wejściowych prognozy. Do głównych komponentów niepewności otrzymywanej prognozy należą: niepewność wynikająca z czynnika losowego zależności między zmiennymi wejściowymi i prognozowanym zapotrzebowaniem; niepewność wynikająca z oszacowania parametrów (wag) modelu na podstawie pewnej skończonej próby danych (zbioru treningowego); propagacja ewentualnej niepewności wejść.

Czynnik losowy (szum losowy) to nieodłączny komponent niepewności prognozy, związany ze stochastyczną zależnością między zapotrzebowaniem na energię elektryczną a zjawiskami, na podstawie których usiłujemy je przewidywać. Model zawsze stanowi przybliżoną reprezentację rzeczywistości, do której wybiera się elementy najważniejsze z punktu widzenia jego tworzenia. Wpływ elementów, których uwzględnienie jest niemożliwe bądź nieracjonalne, manifestuje się w formie czynnika losowego. Analizując badane predyktory, stwierdziliśmy, że w rozważanych zadaniach krótkoterminowej prognozy zapotrzebowania na energię elektryczną mieliśmy do czynienia z szumem losowym o rozkładzie normalnym, ale o zmiennym odchyleniu standardowym, zależnym od wzorca wejściowego prognozy. Model odchylenia standardowego szumu losowego stworzony został w postaci dodatkowej sieci neuronowej.

Szacowanie odchylenia standardowego prognozy, wynikającego z niepewności parametrów modelu (wag sieci), w przypadku modeli nieliniowych jest zadaniem złożonym. Metody stosowane do rozwiązywania tego zadania mają charakter przybliżony i powinny być weryfikowane dla każdego konkretnego zagadnienia prognostycznego. W bieżącym rozdziale przebadaliśmy więc najważniejsze z nich, dla rozważanych przez nas modeli neuronowych i neuronowo-rozmytych, wykorzystując je na obszernym zestawie zadań krótkoterminowego prognozowania zapotrzebowania na energię elektryczną. Otrzymane wyniki wskazują na poprawne działanie badanych metod. W każdym przypadku najlepsze efekty osiągnęliśmy przy wykorzystaniu metody delta z pełnym oszacowaniem hesjanu błędu oraz metody opartej na bootstrapie.

W zagadnieniach krótkoterminowego prognozowania zapotrzebowania na energię elektryczną niepewność wejść związana jest przede wszystkim z wykorzystaniem informacji pogodowej. Tworząc prognozę zapotrzebowania na energię, posługujemy się przewidywaniami dotyczącymi pogody, a więc informacjami niepewnymi, obarczonymi błędem. W rozdziale tym przyjrzeliśmy się kilku metodom uwzględniania niepewności danych wejściowych. W naszych pracach nie udało się uzyskać wyników lepszych niż te wynikające ze standardowego prawa propagacji błędów, zastosowanego do predyktorów neuronowych i neuronowo-rozmytych, jednakże sytuacja może się zmienić przy większej liczbie uwzględnianych w prognozie czynników pogodowych.

Podsumowując, w bieżącym rozdziale pokazaliśmy, że stosowane obecnie metody modelowania niepewności prognozy mogą być wykorzystywane w badanych zadaniach krótkoterminowego zapotrzebowania na energię. Przebadaliśmy szereg modeli neuronowych i neuronowo-rozmytych oraz znaczny zestaw zadań z badanej dziedziny. Otrzymane wyniki potwierdzają hipotezę badawczą postawioną w naszej pracy, a co za tym idzie – stanowią fundamentalny czynnik pozwalający na wykazanie prawdziwości jej tezy.

ROZDZIAŁ 4

Prognozy zapotrzebowania na energię i ryzyko decyzji

W rozdziale 2 zaprezentowaliśmy problematykę wykorzystania sieci neuronowych i neuronowo-rozmytych w zagadnieniach prognozowania krótkoterminowego obciążenia sieci elektroenergetycznej (zapotrzebowania na energię i moc). Obecnie spróbujemy przyjrzeć się kwestiom związanym z wykorzystaniem prognoz zapotrzebowania w procesach podejmowania decyzji na rynku energii. W punkcie 3.1.1 pokazaliśmy, że modele regresyjne, m.in. sieci neuronowe czy neuronowo-rozmyte, prognozują wartość oczekiwaną zapotrzebowania na energię (moc). Zgodnie z tezą naszej pracy, znajomość wyłącznie wartości oczekiwanej prognozowanego popytu z punktu widzenia procesu decyzyjnego w wielu przypadkach może być niewystarczająca.

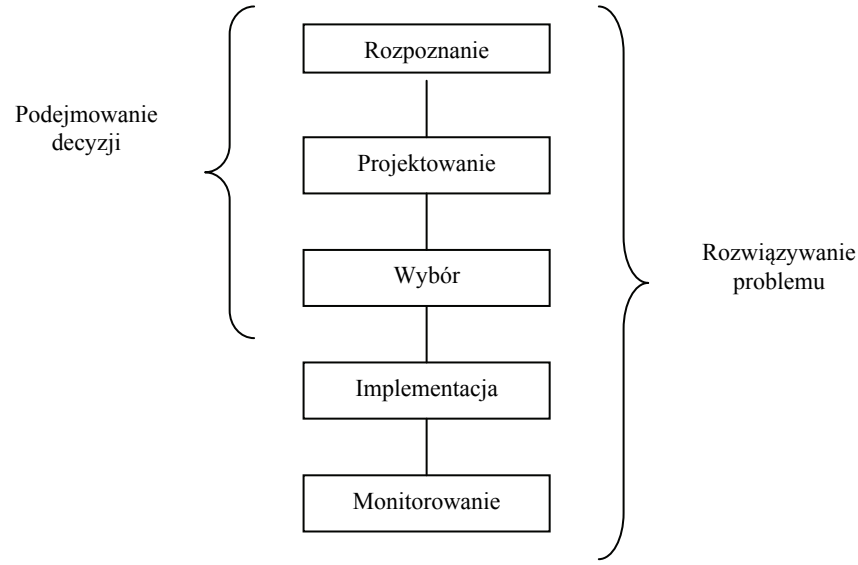
Często bowiem by podjąć optymalną decyzję, trzeba uwzględnić element ryzyka związanego z istniejącą sytuacją i naszą wiedzą o jej istotnych aspektach, a co za tym idzie – przeanalizować niepewności wykorzystywanych przez nas prognoz. W tym celu niezbędne są dodatkowe informacje otrzymywane z całego wynikowego rozkładu prawdopodobieństwa prognozowanej wielkości. Metody określania rozkładu prawdopodobieństwa predykcji dla modeli neuronowych i neuronowo-rozmytych omawialiśmy w rozdziale 3, badając ich zastosowanie w zagadnieniach krótkoterminowego prognozowania zapotrzebowania na energię i wykazując dzięki temu prawdziwość hipotezy badawczej naszej pracy o możliwości zastosowania tych metod w analizowanej dziedzinie.

By ostatecznie potwierdzić tezę postawioną w prezentowanej pracy, w bieżącym rozdziale pokażemy sposób wykorzystania tego rodzaju dodatkowej informacji o rozkładzie prawdopodobieństwa prognozy zapotrzebowania na energię elektryczną w podstawowych typach problemów decyzyjnych, które mogą występować na rynkach energii – w analizie podejmowanych decyzji, szacowaniu ich ryzyka oraz w wyborze optymalnego sposobu postępowania w warunkach tego ryzyka.

4.1. Ogólna charakterystyka procesu podejmowania decyzji

Proces podejmowania decyzji stanowi część, a konkretnie pierwszą fazę, szerszego procesu określanego jako rozwiązywanie problemu. Tradycyjnie (Stair 1992; Jabłoński, Bartkiewicz 2006) rozwiązywanie problemu obejmuje pięć odrębnych etapów zaprezentowanych na rysunku 4.1.1, z których pierwsze trzy wiążą się z samym podjęciem decyzji, natomiast ostatnie dwa – z jej wprowadzeniem w życie oraz obserwacją skutków i efektów podjętych działań. Istotny jest również fakt, że proces ten nie ma charakteru czysto sekwencyjnego. Podczas realizacji kolejnych jego faz może zachodzić potrzeba iteracyjnego powrotu do któregoś z wcześniejszych etapów i powtórnej realizacji wykonywanych podczas niego czynności.

Na etapie rozpoznania identyfikowane i definiowane są problemy, cele, jakie mają zostać osiągnięte i ogólne możliwości ich realizacji. Zbiera się informacje odnoszące się do zakresu problemu i jego otoczenia. Badane są możliwe podejścia do jego rozwiązania, a także ograniczenia, z którymi możemy mieć do czynienia. Z punktu widzenia systemu informacyjnego zarządzania, wspomaganie procesu podejmowania decyzji na etapie rozpoznania polega przede wszystkim na dostarczaniu decydentowi informacji. Wykorzystane mogą być także różnego rodzaju symulacje i komputerowe badania analityczne.



Rysunek 4.1.1. Etapy procesu rozwiązywania problemu i podejmowania decyzji
Źródło: opracowanie własne na podstawie R.M. Stair, *Principles of Information Systems – A Managerial Approach*, Thomson Publishing, Boston 1992

Na etapie projektowania budowane są możliwe rozwiązania analizowanego problemu w formie zestawu (skończonego lub nieskończonego) alternatyw decyzyjnych, które będą dalej rozważane, oraz szacowana jest wykonalność poszczególnych wariantów. Wspomaganie podejmowania decyzji przez system informacyjny zarządzania wiąże się głównie z tworzeniem i kalkulacjami kryteriów oceny poszczególnych rozwiązań alternatywnych oraz z pomocą w odrzucaniu wariantów niewykonalnych. Istotną rolę może odgrywać także asysta przy ocenie elementów ryzyka poszczególnych opcji.

Etap wyboru to ostatnia faza podejmowania decyzji. Wymaga on określenia toku akcji podejmowanych w celu rozwiązania problemu. Ten, na pierwszy rzut oka, łatwy akt wyboru zwykle nie jest taki prosty jak mogłoby się to wydawać. Istnieje wiele czynników, trudnych do rozpoznania, często o charakterze niejawnym, które mogą wpływać na ostateczny wybór alternatywy decyzyjnej. Niezbędne jest wyważenie zarówno stopnia realizacji postawionych celów decyzji, jak i ryzyka z nimi związanego. W przypadku nieskończonej liczby alternatyw decyzyjnych na tym etapie często zastosowanie znajdują komputerowe metody badań operacyjnych, które pozwalają na znalezienie rozwiązania optymalnego pod kątem założonych celów lub przynajmniej określenie najlepszego rozwiązania spośród rozważanych.

Etapy implementacji i monitorowania wykraczają już poza fazę podejmowania decyzji. Polegają one na wykonaniu akcji wprowadzających rozwiązanie (podjętą decyzję) w życie oraz na ocenie implementacji tego rozwiązania przez decydenta, który określa, czy zostały osiągnięte przewidywane efekty i rozważa ewentualną modyfikację całego procesu w świetle nowej informacji zdobytej podczas implementacji. Może to wywołać sprzężenie zwrotne i powrót do któregoś z wcześniejszych etapów rozwiązywania problemu.

Podejmowanie decyzji polega więc na formułowaniu odpowiednich celów, a następnie na określeniu i wyborze sposobu ich realizacji. Oczywiście poszczególne konkretne sytuacje decyzyjne mogą się znacznie różnić między sobą pod względem stopnia złożoności. Niektóre decyzje są stosunkowo proste, przy podejmowaniu innych wybór najlepszego rozwiązania nastręcza olbrzymie trudności. W innych jeszcze przypadkach poważnym wyzwaniem okazuje się nawet samo sformułowanie rozwiązywanego problemu stanowiącego cel podejmowanej decyzji. Z tego punktu widzenia tradycyjnie w przypadku problemów decyzyjnych mówić możemy o następujących kategoriach:

– decyzje strukturalne (programowalne) – łatwe, powtarzalne, rutynowe problemy decyzyjne, dla których istnieją standardowe rozwiązania dające się często przedstawić w postaci określonej procedury postępowania, algorytmu; dla decyzji tego typu modele decyzyjne zostały zbudowane już wcześniej, podjęcie decyzji polega więc jedynie na zastosowaniu znanego rozwiązania; niejednokrotnie decyzje o charakterze strukturalnym podejmowane są automatycznie przez systemy informatyczne,

– decyzje niestrukturalne (nieprogramowalne) – rozmyte, kompleksowe, nierutynowe problemy, dla których nie ma łatwych rozwiązań; nie ma dla nich wypracowanej metodologii ani modelu ich rozwiązania; do ich podejmowania niezbędne są subiektywne sądy i intuicja menedżera,

– decyzje częściowo strukturalne (częściowo programowalne) – w ich przypadku jedynie pewne fazy procesu podejmowania decyzji mają charakter strukturalny; nie mogą one zostać w pełni zautomatyzowane; wymagają subiektywnych ocen i osądów w powiązaniu z formalną analizą danych i budową modeli.

Z punktu widzenia procesu wspomagania decyzji istotny element stanowi również liczba rozważanych alternatyw decyzyjnych. Jeżeli jest ona skończona i niezbyt duża, to wysiłki koncentrują się przede wszystkim na wspomaganiu budowy i oceny poszczególnych alternatyw. Proces samego wyboru optymalnego wariantu jest już zazwyczaj stosunkowo prosty. Ponieważ mamy niezbyt wiele alternatyw decyzyjnych, możemy enumeratywnie wyznaczyć kryterium oceny dla każdego z nich i, w większości przypadków, bez żadnych poważniejszych problemów wybrać najlepsze czy też utworzyć ranking. W przypadku nieskończonej (lub skończonej, ale bardzo dużej) liczby rozwiązań alternatywnych jest to, oczywiście, niemożliwe, proces wyboru wymaga optymalizacji funkcji celu, czyli kryterium oceniającego poszczególne warianty decyzji w warunkach istniejących ograniczeń.

Sytuacja komplikuje się, gdy określając cele podejmowanej decyzji, musimy wziąć pod uwagę kilka kryteriów oceny alternatyw. Poszczególne kryteria zazwyczaj nie są w pełni zgodne między sobą (gdyby były, to prawdopodobnie nie musielibyśmy rozważać kilku kryteriów), a czasami nawet bywają wręcz sprzeczne. Na przykład podejmując decyzję o zakupie jakiegoś urządzenia, kierujemy się zazwyczaj przeciwstawnymi celami, jakimi mogą być cena, parametry jakościowe czy też koszty eksploatacyjne rozważanych konkretnych modeli danego urządzenia. W związku z tym niezbędne okazuje się scalenie (agregacja) poszczególnych kryteriów w sztuczną miarę oceny alternatyw decyzyjnych nazywaną zwykle użytecznością. Wymaga to zazwyczaj dostarczenia przez decydenta informacji o istotności (lub preferencji) poszczególnych kryteriów lub analizy zbioru danych historycznych.

Innym ważnym czynnikiem wpływającym na decydenta podczas analizy możliwych wariantów postępowania są rozmiary ryzyka związanego z daną decyzją. Pojęcie ryzyka ma naturalnie bardzo szeroki charakter i może być definiowane oraz interpretowane w rozmaity sposób. W bieżącym rozdziale, jeżeli nie zostanie to wyraźnie powiedziane inaczej, będziemy je rozumieli przede wszystkim w sensie probabilistycznym. Ryzykiem będziemy więc określali pewną ocenę prawdopodobieństwa faktu, że wybrane przez decydenta rozwiązanie problemu może zaowocować nieoczekiwanym lub niepożądanym wynikiem.

Z tego punktu widzenia możemy więc mówić o następujących rodzajach decyzji:

– decyzje podejmowane w warunkach pewności – występują w sytuacjach, gdy decydent w fazie rozpoznania zebrał informacje na tyle wyczerpujące i dokładne, aby móc dogłębnie zrozumieć rozwiązywany problem; wszystkie istotne fakty mogące wpłynąć na wynik decyzji znane są decydentowi tak, że jest on w stanie z góry określić wynik każdej alternatywy; nie ma więc ryzyka, że rozwiązanie przyniesie nieoczekiwane efekty; oczywiście w praktyce znajomość absolutnie wszystkich faktów jest w zasadzie niemożliwa; tym niemniej w pewnych sytuacjach decydent może wiedzieć na tyle dużo, aby czynnik ryzyka można było pominąć; z tego typu zjawiskiem na ogół mamy do czynienia w przypadku decyzji strukturalnych (oraz do pewnego stopnia częściowo strukturalnych),

– decyzje podejmowane w warunkach niepewności – występują w sytuacjach, gdy system informacyjny o pewnych faktach lub zjawiskach nie jest w stanie dostarczyć decydentowi żadnej wiedzy; jedynym rozwiązaniem w takim przypadku jest stosowanie technik symulacyjnych i analizowanie efektów decyzji dla różnych wartości nieznanych zmiennych,

– decyzje podejmowane w warunkach ryzyka – wiążą się z rozwiązywaniem problemów, dla których możemy określić jedynie prawdopodobieństwo wyniku; w rzeczywistości w działalności gospodarczej niewiele jest decyzji, dla których wszystkie istotne fakty znane są w momencie ich podejmowania; przy tych decyzjach decydent nie wie z całą pewnością, czy wybrane rozwiązanie da oczekiwane efekty, stąd pojawia się ryzyko.

Jeszcze raz zwróćmy uwagę na fakt, że podejmowanie decyzji nieodłącznie wiąże się z towarzyszącym temu procesowi ryzykiem. W zasadzie – poza stosunkowo prostymi zagadnieniami – decydent niemal zawsze stoi przed koniecznością wyboru dalszego sposobu postępowania, nie mając pełnej wiedzy o ważnych zmiennych decyzyjnych i parametrach, które wywierają wpływ na ocenę rozważanych alternatyw i stopień realizacji założonych celów. Stąd właśnie istotna rola w procesie decyzyjnym prognozowania, które pozwala na dostarczenie informacji oraz oszacowań niezbędnych do redukcji czynników ryzyka występujących w danej sytuacji decyzyjnej.

Prognozowanie w procesie podejmowania decyzji pełni naturalnie funkcje pomocnicze. Nie jest ono celem samym w sobie, lecz istotnym narzędziem, które służy do redukcji niepewności decydenta co do istotnych zjawisk wpływających na podejmowane decyzje. Aby jednak mogło ono odgrywać tę rolę, niezbędne jest pełne zrozumienie wpływu prognozy na ocenę rozważanych alternatyw oraz identyfikacja wszystkich źródeł niepewności z nią związanych (Marshall, Oliver 1995). Jak już kilkakrotnie zwracaliśmy uwagę, prognoza nie jest informacją nieomylną i pewną. W celu oszacowania ryzyka podejmowanych

decyzji należy więc koniecznie przeanalizować i ocenić niepewność prognozy (Bartkiewicz 2002).

Każda prognoza nieodłącznie obarczona jest błędem, wykorzystując ją więc w problemach decyzyjnych decydent musi być świadomy niepewności, jaka wiąże się z wykorzystywanym oszacowaniem oraz wynikającym z tej niepewności ryzykiem dla podejmowanej decyzji. Ocena ta dokonywana musi być w sposób zgodny z zasilanym przez prognozę problemem decyzyjnym. Wykorzystanie w tym celu kryteriów jakości prognozy opartych wyłącznie na statystycznych miarach dopasowania modelu prognostycznego do danych treningowych lub testowych powoduje, że często szacowane są niewłaściwe aspekty problemu prognostycznego. Kryteria te nie są bowiem bezpośrednio powiązane z procesami decyzyjnymi zasilanymi przez system prognostyczny.

Zarówno sam proces prognozowania, jak i ocena osiągniętych wyników nie mogą zatem odbywać się w oderwaniu od podejmowanych na podstawie prognozy decyzji. Możemy tutaj wręcz zacytować za Kneale T. Marshallem i Robertem M. Oliverem:

> Do najważniejszych wymagań należy, by gromadzenie danych, prognozowanie i podejmowanie decyzji stanowiły spójne przedsięwzięcie i miały zgodne ze sobą cele. Projektant(ci) baz danych, modelu(i) prognostycznych i decydent(ci) powinni myśleć w jednolity sposób. Nawet jeżeli jest inaczej, jak to się łatwo może zdarzyć, gdy dane i prognozy otrzymywane są z różnych źródeł zewnętrznych, decydent musi dobrze rozumieć kwestię wyjść modelu prognostycznego oraz ich jakości. Fakt, że prognozy muszą być zgodne z potrzebami modelu decyzyjnego stanowi bardzo ważną sprawę (Marshall, Oliver 1995, s. 25, tłum. moje – W.B.).

4.2. Prognozy i decyzje

Jak podkreślaliśmy w poprzednim punkcie, jeden z najważniejszych priorytetów polega na zapewnieniu spójności procesu prognozowania i podejmowania decyzji. Minimalnym warunkiem, jaki musimy tutaj przyjąć, jest wymaganie, aby decydent otrzymywał z modelu prognostycznego wszelkie istotne informacje, które służą do rozważenia różnych aspektów podejmowanej decyzji. W odniesieniu do omawianych w naszej pracy modeli neuronowych i neuronowo-rozmytych często istotny element stanowić więc będzie informacja o rozkładzie prawdopodobieństwa prognozy. Dzięki niej bowiem decydent jest w stanie rozważyć niepewność otrzymywanego oszacowania i wykorzystać te informacje w konkretnym problemie decyzyjnym.

W rozdziale 3 omawialiśmy metody wnioskowania na temat rozkładu prognozy dla nieliniowych punktowych predyktorów regresyjnych. Obecnie

zajmiemy się wykorzystaniem uzyskanych w ten sposób informacji w problemach analizy decyzyjnej. W bieżącym podrozdziale ograniczymy się do sytuacji skończonego zbioru alternatyw decyzyjnych, a naszym celem będzie omówienie oraz ilustracja kilku istotnych zagadnień pojawiających się na styku prognozy, podejmowanych na ich podstawie decyzji oraz ryzyka z tym związanego. Zagadnieniami nieskończonej liczby alternatyw decyzyjnych zajmiemy się w następnych podrozdziałach, zastanawiając się zarazem nad problemami określania optymalnej wielkości zakupów i sprzedaży.

4.2.1. Prognozy zapotrzebowania na energię jako dyskretne zmienne losowe

Przyjrzyjmy się obecnie zagadnieniom decyzyjnym, w których wynik określonej alternatywy zależny jest od dyskretnej zmiennej popytowej obarczonej niepewnością. Rozważymy to zadanie na przykładzie prostego problemu, w którym nasza decyzja zależy od wielkości zapotrzebowania na energię.

Problem 4.2.1

Rozważmy pewną inwestycję, której wynik zależy od zapotrzebowania na energię elektryczną lub moc w krótkiej perspektywie czasowej. Może chodzić tutaj o zakup kontraktu krótkoterminowego, działania związane z zadaniami zarządzania stroną popytową (*Demand Side Management*, DSM) (Belina, Węgliński, Zieliński 1996) czy też z tworzeniem rezerwy mocy w systemie lokalnym. Liczymy przy tym na to, że zapotrzebowanie nie przekroczy pewnej granicy g. Dokładniej mówiąc, szacujemy, że w przypadku sukcesu, czyli jeżeli zapotrzebowanie nie będzie wyższe niż g, osiągniemy zysk o wartości r_s. W przypadku porażki, czyli jeżeli zapotrzebowanie przekroczy tę granicę, nasz zysk będzie znacznie niższy (być może będziemy mieli do czynienia nawet z wartością ujemną – stratą) r_p. Rozwiązaniem alternatywnym jest decyzja bezpieczna, której efekty nie zależą od zapotrzebowania – daje nam ona zysk r_b. Zakładamy przy tym, rzecz jasna, że najgorsze efekty uzyskujemy w przypadku porażki, najlepsze – w przypadku sukcesu, czyli $r_p < r_b < r_s$. Niestety w chwili podejmowania decyzji nie dysponujemy żadną wiedzą na temat możliwej wielkości zapotrzebowania na energię, z jaką będziemy mieli do czynienia.

Dokładne wartości wypłat dla poszczególnych alternatyw decyzyjnych mają w tej chwili mniejsze znaczenie, ale przyjmijmy dla przykładu, że zysk w przypadku sukcesu wynosi r_s = 25 000 złotych, w przypadku porażki tracimy znacznie więcej r_p = –150 000 złotych, zaś bezpieczna kwota r_b = 10 000 złotych. Natomiast wartość granicznego zapotrzebowania na energię, poniżej którego odniesiemy sukces, równa jest g = 200 000 kWh.

Widzimy więc, że w problemie 4.2.1 musimy wybrać spośród dwóch alternatyw decyzyjnych: bezpiecznej, którą oznaczmy przez *AB*, oraz obciążonej ryzykiem – *AR*. Jako racjonalne kryterium ekonomiczne podjęcia decyzji przyjmiemy chęć maksymalizacji zysku. Ryzyko pojawia się więc z powodu naszej niepewności co do wyniku decyzji w przypadku wyboru *AR*. Nie znamy przecież faktycznego stanu rzeczywistości związanego z prawdziwą realizacją prognozowanego poziomu zapotrzebowania na energię. Jeżeli rzeczywiste zapotrzebowanie utrzyma się w granicach *g*, zarobimy więcej niż w przypadku wyboru bezpiecznej alternatywy decyzyjnej *AB*, w przeciwnym przypadku – mniej.

W sytuacji, w której nie dysponujemy żadnymi prognozami zapotrzebowania na energię, nie mamy również żadnego sposobu redukcji naszej niepewności związanej z wynikiem ryzykownej decyzji *AR*. Nie możemy przewidzieć, który z możliwych stanów rzeczywistości zrealizuje się faktycznie: sukces czy porażka. Jedynym więc sposobem pozwalającym na uporządkowany wybór lepszego rozwiązania jest zastosowanie metod analizy decyzji w warunkach niepewności. Istnieje kilka podejść, które możemy tutaj wykorzystać. Polegają one na przebadaniu wszystkich możliwych skutków wyboru poszczególnych alternatyw dla wszystkich możliwych stanów rzeczywistości. Następnie należy uporządkować te rozwiązania przy użyciu pewnych przyjętych kryteriów ostatecznego wyboru, zazwyczaj takich jak zyski, koszty lub wielkości od nich pochodne.

Zestawienie wyników, które możemy osiągnąć w przypadku wyboru alternatywy ryzykownej *AR* oraz bezpiecznej *AB* dla naszego problemu 4.2.1, znajduje się w tabeli 4.2.1.

Tabela 4.2.1. Tabela możliwych wyników poszczególnych alternatyw decyzyjnych dla problemu 4.2.1

Zapotrzebowanie na energię (*ZE*)	Alternatywa ryzykowna (*AR*)	Alternatywa bezpieczna (*AB*)
Sukces ($ZE \leq g$)	r_s	r_b
Porażka ($ZE > g$)	r_p	r_b

Źródło: opracowanie własne.

Najbardziej pesymistycznym podejściem do rozwiązania problemu podjęcia decyzji w warunkach niepewności jest reguła max-min (tzw. reguła Walda). Polega ona na wyborze alternatywy decyzyjnej, która w najgorszym przypadku oferuje najlepszy wynik. Odpowiada więc ona regule podejmowania decyzji opartej przede wszystkim na bezpieczeństwie, czyli minimalizacji ryzyka ewentualnych niekorzystnych skutków wybranego rozwiązania. W przypadku

ogólnym kryterium max-min możemy zapisać w następujący sposób. Jeżeli przez r_{ij} oznaczymy wynik i-tej alternatywy decyzyjnej dla j-tego stanu rzeczywistości wpływającego na jej skutki, to kryterium oceny tej alternatywy jest miara $kr(i)$, zdefiniowana następująco (spodziewamy się najgorszego):

$$kr(i) = \min_j r_{ij} \tag{4.2.1}$$

Wybieramy oczywiście alternatywę decyzyjną, dla której wartość kryterium $kr(i)$ jest największa. Jeżeli spojrzymy do tabeli 4.2.1, widzimy od razu, że alternatywa ryzykowna AR w najgorszym przypadku daje wynik r_p. Alternatywa bezpieczna AB daje stały wynik o wysokości r_b. Ponieważ zakładaliśmy, że $r_p < r_b$, więc podejmujemy decyzję bezpieczną, która w tym najgorszym przypadku pozwala osiągnąć najlepsze wyniki.

Innym możliwym kryterium wyboru, leżącym niejako na przeciwległym biegunie, jest reguła optymistyczna max-max. Tym razem wybieramy alternatywę decyzyjną oferującą najlepsze wyniki w najlepszym przypadku, co odpowiada regule podejmowania decyzji, dzięki której potencjalnie osiąga się najwięcej, nie zważając na ryzyko, jakie się z nią wiąże. Kryterium oceny poszczególnych rozważanych alternatyw decyzyjnych $kr(i)$ przyjmuje więc obecnie postać:

$$kr(i) = \max_j r_{ij} \tag{4.2.2}$$

gdzie, podobnie jak w poprzednim przypadku, przez r_{ij} oznaczymy wynik i-tej alternatywy decyzyjnej dla j-tego stanu rzeczywistości.

Tak samo zatem wybieramy alternatywę decyzyjną, dla której wartość kryterium $kr(i)$ jest największa. I znów, jeśli spojrzymy do tabeli 4.2.1, to widzimy, że najlepszy wynik, jaki możemy osiągnąć dla alternatywy ryzykownej AR, wynosi r_s. W przypadku bezpiecznego rozwiązania, podobnie jak dla pesymistycznej reguły wyboru, osiągamy stały wynik w wysokości r_b. Ponieważ więc $r_b < r_s$, powinniśmy wybrać alternatywę ryzykowną jako tę, która pozwoli na osiągnięcie potencjalnie lepszych wyników.

Przedstawione kryteria stanowią naturalnie pewne ekstremalnie przeciwstawne przypadki pod względem podejścia decydenta do ryzyka związanego z podejmowaną decyzją. Strategię pośrednią, pozwalającą na modelowanie czynnika skłonności do ryzyka, stanowi metoda Hurwicza. Umożliwia ona wypracowanie kryterium wyboru alternatywy decyzyjnej, które okaże się kompromisem między postawą pesymistyczną a optymistyczną. Jako podstawę decyzji w regule Hurwicza wykorzystuje się średnią ważoną z najlepszego i najgorszego wyniku każdego rozważanego wariantu decyzji. W tym przypadku

kryterium oceny poszczególnych alternatyw decyzyjnych $kr(i)$ możemy więc zapisać jako:

$$kr(i) = \lambda \max_j r_{ij} + (1-\lambda) \min_j r_{ij} \tag{4.2.3}$$

gdzie λ jest stałą z przedziału [0, 1] określającą zakładany poziom optymizmu (skłonność do ryzyka) decydenta. Ponieważ w tym przypadku również wybierając rozwiązanie, maksymalizujemy wartość kryterium $kr(i)$, widzimy, że im większa wartość λ, tym bardziej wynik reguły Hurwicza zbliża się do optymistycznej reguły max-max, by w końcu dla $\lambda = 1$ zredukować się do tego kryterium. Przeciwnie, jeżeli wartość λ maleje do 0, reguła sprowadza się do zachowawczego kryterium max-min. W przypadku problemu 4.2.1 alternatywę ryzykowną AR wybierzemy, gdy dla założonego poziomu optymizmu λ spełniony będzie warunek:

$$\lambda r_s + (1-\lambda) r_p > r_b \tag{4.2.4}$$

Jeżeli więc przyjmiemy współczynnik optymizmu $\lambda = 0{,}95$, to przy danych z problemu 4.2.1 widzimy, że:

– dla alternatywy ryzykownej AR: $kr(AR) = 0{,}95 \cdot 25\,000 + 0{,}05 \cdot (-150\,000) = 16\,250$,

– w przypadku alternatywy bezpiecznej AB: $kr(AR) = 10\,000$.

W tym przypadku, ponieważ decydent jest dosyć optymistyczny, powinien wybrać rozwiązanie ryzykowne. Problem polega na tym, jak określić poziom naszego optymizmu w konkretnej sytuacji. Pamiętajmy, że nie potrafimy powiedzieć niczego o tym, który wariant rzeczywistości faktycznie się zrealizuje, więc jakikolwiek wybór współczynnika λ będzie zależny od czynników subiektywnych, niewynikających z obiektywnej wiedzy na temat decyzji – takich jak nastawienie psychologiczne decydenta.

Kolejną regułą porządkowania alternatyw decyzyjnych w warunkach niepewności, opartą na nieco innej zasadzie, jest tzw. reguła utraconej szansy Niehansa–Savage'a. U jej podstaw leży idea sformułowania kryterium wyboru wariantu decyzji nie w postaci jego bezpośrednich skutków, ale w postaci odniesienia się do straty spowodowanej tym, że nie wybraliśmy najlepszej alternatywy decyzyjnej dla każdego stanu rzeczywistości. Decydenci bowiem często obserwując efekty decyzji po wprowadzeniu jej w życie, porównują to, co otrzymują, z tym, co mogliby otrzymać, poprawnie identyfikując realizację stanu rzeczywistości. Obserwowaną różnicę odbierają oni jako swoistą stratę albo „utraconą szansę". Należy więc wybrać alternatywę decyzyjną, która taką ewentualną utraconą szansę zminimalizuje. Sposób postępowania możemy zreasumować tu następująco. Po pierwsze, dla każdego stanu rzeczywistości

znajdujemy najlepszy wynik ze wszystkich możliwych wariantów decyzji, które mamy do wyboru. Potem wyznaczamy różnicę między tym optimum a wynikiem każdego alternatywnego rozwiązania, czyli wielkość straty z naszej utraconej szansy. Oznaczmy tę stratę dla *j*-tego stanu rzeczywistości oraz *i*-tej alternatywy decyzyjnej przez us_{ij}. I tak:

$$us_{ij} = \max_i r_{ij} - r_{ij} \tag{4.2.5}$$

Następnie jako kryterium wyboru bierzemy pod uwagę maksymalną szansę, jaką moglibyśmy dla tej alternatywy utracić:

$$kr(i) = \max_j us_{ij} \tag{4.2.6}$$

Tabela 4.2.2. Tabela utraconych szans poszczególnych alternatyw decyzyjnych dla problemu 4.2.1

Zapotrzebowanie na energię (*ZE*)	Alternatywa ryzykowna (*AR*)	Alternatywa bezpieczna (*AB*)
Sukces ($ZE \leq g$)	0	$r_s - r_b$
Porażka ($ZE > g$)	$r_b - r_p$	0

Źródło: opracowanie własne.

Oczywiście tym razem wybierzemy alternatywę minimalizującą wartość naszego kryterium, czyli minimalizującą maksymalną stratę z ewentualnie utraconej szansy. W przypadku problemu 4.2.1 sprowadza się to do dosyć prostych kalkulacji. W tabeli 4.2.2 przedstawione zostały wielkości utraconych szans dla obu alternatyw decyzyjnych, przy różnych stanach rzeczywistości. Widzimy więc, że jeżeli zapotrzebowanie na energię będzie poniżej założonego progu, to dla decyzji ryzykownej *AR* nie ma żadnej utraconej szansy, a ewentualna strata wynosi 0. Jeżeli zapotrzebowanie będzie zbyt duże, to utracona szansa wynosi $r_b - r_p$ (jest to wartość większa od 0). Analogicznie dla decyzji bezpiecznej *AB* utracone szanse wynoszą $r_s - r_b$ (co również jest wartością dodatnią) lub 0. Bez trudu otrzymujemy więc regułę wyboru: jeżeli ewentualny zysk z sukcesu (mierzony w stosunku do alternatywy bezpiecznej) jest większy niż ewentualna strata z porażki, decydent powinien wybrać rozwiązanie ryzykowne. W przeciwnym razie należy zdecydować się na decyzję bezpieczną. Dla przykładowych danych mamy więc:

– dla alternatywy ryzykownej *AR*: $kr(AR) = 10\,000 - (-150\,000) = 160\,000$,

– w przypadku alternatywy bezpiecznej *AB*: $kr(AR) = 25\,000 - 10\,000 = 15\,000$.

W tym przypadku widzimy więc, że decydent powinien wybrać wyjście bezpieczne jako to, które minimalizuje ewentualną stratę z niewykorzystanych szans.

Zaprezentowaliśmy kilka możliwości w zakresie porządkowania i wyboru alternatyw decyzyjnych w warunkach niepewności. Zwróćmy uwagę, że podstawowym wyróżnikiem poszczególnych metod jest podejście do traktowania ryzyka decyzji w świetle pojawiającej się niepewności odnośnie do istotnych faktów i zjawisk. Niektóre z nich są bardziej zachowawcze, inne bardziej agresywne, co pozwala decydentom na różną reakcję, w zależności od ich optymizmu i skłonności do ryzyka. Problem polega oczywiście na tym, że nie mamy żadnej możliwości oszacowania wielkości czynnika ryzyka dla określonych warunków, dla konkretnej sytuacji decyzyjnej, którą rozważamy w danej chwili. Przypomnijmy przecież, że w problemie 4.2.1 zakładamy, że nie wiemy nic o możliwej wielkości zapotrzebowania na energię elektryczną, a zatem nie potrafimy powiedzieć, czy przekroczy ono założony poziom, czy nie.

W tym właśnie należy upatrywać istotną rolę prognoz, oszacowań czy ekspertyz. Jedynie dzięki nim możemy zredukować naszą niepewność odnośnie do nieznanych faktów i zjawisk wpływających na wynik rozważanych alternatyw decyzyjnych. To one właśnie umożliwiają bardziej dogłębne rozważenie sytuacji, analizę czynników ryzyka, jakie się z nią wiąże, co podnosi jakość podejmowanych decyzji.

Mówiąc o prognozach, myślimy o szerokim spektrum działań opartych na naukowym przewidywaniu przyszłości. Do tej kategorii zaliczyć należy m.in. regresyjne prognozy punktowe w formie oszacowań wartości oczekiwanych (przeciętnych) rozmaitych zmiennych, istotnych dla podejmowanej decyzji. Możemy również mówić o prognozach o charakterze kategorycznym, w których przewiduje się wystąpienie jednego ze skończonej liczby dyskretnych stanów (kategorii wartości) analizowanego zjawiska. Nieco odmienny charakter mają prognozy probabilistyczne w formie stwierdzeń określających prawdopodobieństwo wystąpienia (lub nie) jakiegoś przewidywanego zdarzenia. Przez prognozy rozumiemy oszacowania otrzymywane zarówno za pomocą modeli statystycznej analizy danych historycznych, jak i na podstawie analizy heurystycznej, opartej na wiedzy i doświadczeniu eksperta bądź operatora istotnego procesu, wyrażane w formie subiektywnych opinii i sądów.

Temat powiązań między prognozami i decyzjami ma naturalnie bardzo szeroki charakter, wykraczający znacznie poza zakres naszej pracy. Zainteresowanych Czytelników odsyłamy np. do klasycznej pozycji literatury z tej problematyki (Marshall, Oliver 1995). Zagadnienia te interesować nas będą przede wszystkim jako ilustracja przydatności informacji probabilistycznej uzyskanej za pomocą metod prezentowanych w poprzednim rozdziale – pozwala ona na ocenę niepewności neuronowych i neuronowo-rozmytych prognoz zapotrzebowania na

energię (moc) oraz wynikającego z nich ryzyka decyzji gospodarczych na rynkach energii.

Jak to analizowaliśmy w rozdziale 3, predyktory w postaci sieci neuronowych czy neuronowo-rozmytych, dopasowywane do danych treningowych poprzez minimalizację błędu kwadratowego, szacują zapotrzebowanie na energię w postaci punktowej prognozy wartości oczekiwanej tego procesu. Problem polega na tym, że z punktu widzenia właściwej oceny ryzyka decyzji i analizy możliwych wariantów (alternatyw decyzyjnych) tego typu prognozy mają istotne, ale jednak ograniczone znaczenie. Jest to pogląd dosyć powszechnie reprezentowany przez specjalistów z dziedziny analizy decyzyjnej (patrz np. Marshall, Oliver 1995). Oparcie się na oszacowaniu wyłącznie wartości oczekiwanej, bez uwzględnienia niepewności tego oszacowania, nie pozwala w zasadzie na jakąkolwiek analizę ryzyka związanego ze skutkami decyzji i uwzględnienia elementów tego ryzyka w procesie wyboru najlepszego sposobu postępowania. Należy ponadto pamiętać, że funkcja kwadratowa błędu, powszechnie wykorzystywana w budowie tego typu modeli, ma charakter optymalny z punktu widzenia statystycznego, niekoniecznie zaś z punktu widzenia skutków podejmowanej decyzji.

W tym właśnie należy upatrywać istotną rolę badanych w rozdziale 3 naszej pracy metod rekonstrukcji rozkładu prawdopodobieństwa prognozy dla predyktorów neuronowych i neuronowo-rozmytych. Uwzględnienie tej dodatkowej informacji pozwala na ocenę niepewności uzyskiwanych danych, wykorzystywanych dalej w procesie podejmowania decyzji, z punktu widzenia tej właśnie decyzji, a nie standardowych miar błędów statystycznych. Dzięki temu możemy we właściwy sposób przeanalizować ryzyko rozważanych alternatyw decyzyjnych, dokonując wyboru właściwego wariantu. Przyjrzyjmy się temu zagadnieniu na podstawie sytuacji decyzyjnej stanowiącej rozszerzenie problemu 4.2.1.

Problem 4.2.2

Przyjmijmy, że w przypadku decyzji opisanej w problemie 4.2.1 decydent chciał zredukować swoją niepewność odnośnie do wielkości zapotrzebowania na energię elektryczną stanowiącego podstawę wyboru postępowania. Wykorzystał w tym celu model prognostyczny oparty na sieci neuronowej lub neuronowo-rozmytej, uzyskując prognozę zapotrzebowania w rozważanym okresie, o wartości $f(\mathbf{x}, \mathbf{w})$ = 189 000 kWh, gdzie przez $\mathbf{x}$ rozumiemy wzorzec danych wejściowych prognozy, natomiast przez $\mathbf{w}$ – zestaw wag sieci. Odchylenie standardowe dla tej konkretnej prognozy, uzyskane jedną z metod analizowanych w rozdziale 3, zostało oszacowane na $\sigma(\mathbf{x})$ = 10 000 kWh.

Decydenci mają tendencję do niezauważania ostatniego zdania, które znalazło się w opisie problemu 4.2.2 i kierowania się wyłącznie prognozą analizowanej zmiennej, określoną w zdaniu poprzednim. Sprawa wydaje się dosyć oczywista. Ponieważ prognoza $f(\mathbf{x}, \mathbf{w})$ równa jest 189 000 kWh, zaś próg

zapotrzebowania na energię *ZE* dla naszej decyzji określony w problemie 4.2.1 wynosi g = 200 000 kWh, to w związku z tym mamy sytuację, że spodziewane zapotrzebowanie na energię $f(\mathbf{x}, \mathbf{w})$ kształtuje się na poziomie niższym od g, czyli prognoza jest korzystna. Podejmujemy więc decyzję o inwestycji, wybierając alternatywę ryzykowną *AR*, i spokojnie czekamy na wypłatę należnych nam zysków w wysokości 25 000 złotych.

Taki sposób postępowania nie zawsze jest niewłaściwy. Niestety pomijamy przy nim pewne istotne czynniki wpływające na wybór najlepszej decyzji. Problem polega na tym, że działając w opisany sposób, traktujemy prognozę jako fakt niewątpliwy i pewny. A tak, rzecz jasna, wcale nie jest. Każda prognoza musi być obarczona błędem. Nasza prognoza określa jedynie wartość oczekiwaną zapotrzebowania na energię ($E(ZE) = f(\mathbf{x}, \mathbf{w})$). Faktyczna realizacja tego procesu kształtować powinna się w pobliżu prognozy, ale niemal nigdy nie będzie jej dokładnie równa. Zapotrzebowanie na energię nadal pozostaje zmienną losową. Prognoza pozwala jedynie określić (lub sprecyzować) jego rozkład prawdopodobieństwa. Jeżeli nie weźmiemy tego faktu pod uwagę, możemy po wyborze alternatywy ryzykownej *AR* stanąć przed faktem straty 150 000 złotych, ponieważ zapotrzebowanie na energię elektryczną przekroczy jednak założony przez nas jako bezpieczny poziom g.

Przyjmijmy więc teraz, że nasz hipotetyczny decydent zdaje sobie sprawę, że prognoza jest tylko prognozą, i tak do końca nie może jej ufać. Tak jak zresztą wspominaliśmy, podejmowanie decyzji na podstawie prognozy wartości oczekiwanej wykorzystywanych zmiennych nie zawsze jest działaniem niewłaściwym. W naszym przypadku jeżeli prognoza zapotrzebowania na energię $f(\mathbf{x}, \mathbf{w})$ jest niższa od założonego poziomu g, to możemy podjąć decyzję o wyborze rozwiązania ryzykownego *AR* – pod warunkiem, że różnica między tymi wielkościami jest bezpiecznie duża i ryzyko przekroczenia tej granicy jest na tyle małe, że możemy je w praktyce zaniedbać. Innymi słowy, zakładamy, że prognoza w pełni redukuje naszą niepewność do zera, i kierując się jej wskazaniami, możemy traktować ją jako daną pewną, a w konsekwencji zastosować reguły podejmowania decyzji w warunkach pewności. Sytuację taką określa się czasami jako „decyzję stabilną" w świetle niepewności prognozy.

Nawet w takich warunkach łatwo jednak wpaść w pewien niebezpieczny schemat myślenia. Przyzwyczajeni jesteśmy bowiem do oceny dokładności prognozy w kategoriach pewnych standardowych miar błędów statystycznych. Jako miarę spodziewanego odchylenia prognozy moglibyśmy przyjąć np. średni błąd bezwzględny (MAE) czy też pierwiastek średniego błędu kwadratowego (RMSE). Są to informacje, którymi często dysponujemy, ponieważ twórcy modeli prognostycznych zazwyczaj oceniają dokładność prognozy właśnie za pomocą tego typu mierników. Chętnie zresztą udzielają tych informacji, traktując je jako elementy specyfikacji parametrów sprzedawanego nam, czy też sporządzanego dla nas, systemu.

W problemie 4.2.2 mamy informację, że odchylenie standardowe prognozy zapotrzebowania na energię elektryczną wynosi 10 000 kWh, więc prawdopodobnie na zbliżonym poziomie będą kształtowały się również oceny średnich odchyleń predykcji (zwróćmy uwagę na słowo „zbliżonym"). Przyjmijmy, że tak rzeczywiście jest, czyli błąd tej prognozy, powiedzmy RMSE, faktycznie wynosi około 10 000 kWh. Wydaje się więc, że sytuacja w naszym przykładzie jest dosyć jasna. Różnica pomiędzy prognozą zapotrzebowania na energię a zakładanym progiem wynosi 11 000 kWh, więc jest większa od spodziewanego oszacowania odchylenia prognozy. Nasz hipotetyczny decydent mógłby więc uznać, że spokojnie może podjąć decyzję o zainwestowaniu, czyli o wyborze ryzykownej alternatywy decyzyjnej *AR*.

Podejmując w opisanych warunkach decyzję o inwestycji, nasz decydent nie bierze jednak pod uwagę pewnych istotnych czynników, które mogą wpływać na wyniki jego sposobu rozumowania:

– błąd prognozy jest pewną zmienną losową; zapotrzebowanie na energię to proces stochastyczny; nie potrafimy objaśnić w pełni jego zachowania i nie wiemy dlaczego (nawet w podobnych warunkach) jest ono nieco mniejsze lub większe; błędy średnie typu MAE, RMSE są miernikami średniego odchylenia; wcale nie mamy gwarancji, że dla bieżącego przypadku odchylenie ukształtuje się poniżej średniej,

– tak jak podkreślaliśmy w punkcie 3.2.3, nasza niepewność co do wartości prognozy dla nieliniowego modelu neuronowego lub neuronowo-rozmytego jest proporcjonalna do rozkładu wzorców danych w zbiorze treningowym; innymi słowy, niepewność ta zależna jest od wartości wejściowych konkretnej prognozy i tym mniejsza, im więcej danych treningowych znajdowało się w pobliżu zestawu danych wejściowych tej prognozy; efekt ten uwzględniany jest przez metody szacowania odchylenia standardowego prognozy omawiane w rozdziale 3; standardowe błędy statystyczne oceniają działanie modelu prognostycznego w sposób syntetyczny, średni dla całej przestrzeni wejść; w przypadkach konkretnych odchylenie standardowe prognozy i syntetyczne miary średniego odchylenia mogą się poważnie różnić,

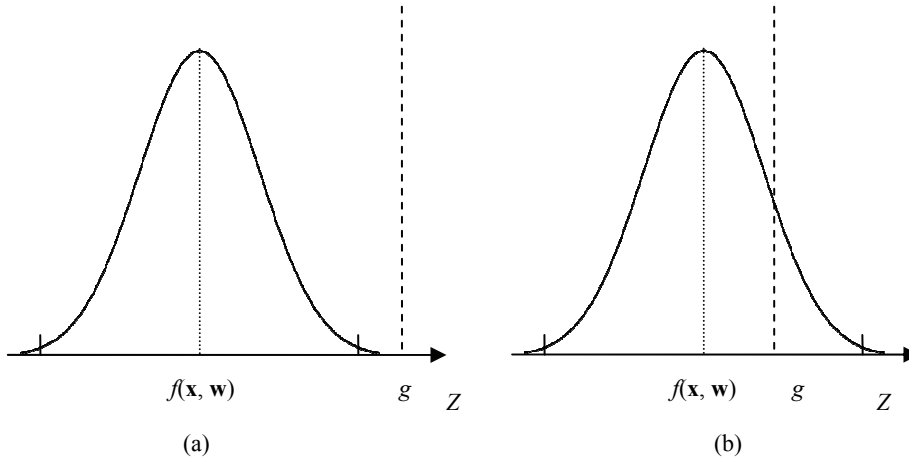
– standardowe miary błędu statystycznego służą do obiektywnej oceny prognozy, niezależnie od sposobu jej wykorzystania; pełnią one istotną rolę w procesie tworzenia modelu czy też porównywania różnych modeli; nie biorą jednak pod uwagę konkretnej subiektywnej sytuacji decyzyjnej, wielkości zysków/strat związanych z rozważanymi alternatywami decyzyjnymi; mogą więc być mało adekwatne w ocenie ryzyka związanego z daną decyzją.

Właściwym sposobem postępowania jest więc wykorzystanie rozkładu prawdopodobieństwa prognozy i rozważenie go w kontekście analizy skutków podejmowanej decyzji. Do wstępnej oceny stabilności decyzji oraz naszej pewności co do jej możliwych skutków możemy dla przykładu wykorzystać przedziały prognozy. Zagadnienie ich wyznaczania szeroko omawialiśmy

w wielu punktach rozdziału 3. Przypomnijmy, że przez przedział prognozy z prawdopodobieństwem α określamy taki przedział wartości prognozowanej zmiennej, w którym znajdować się ona powinna z prawdopodobieństwem α. Informacja tego typu pozwala więc decydentowi na ustanowienie pewnych ram wartości, w których może znajdować się interesująca go zmienna, co w naturalny sposób posłużyć może do analizy bezpieczeństwa podejmowanej decyzji.

Wróćmy więc do naszego przykładu 4.2.2. Omawiając w rozdziale 3 problem określania rozkładu prawdopodobieństwa krótkoterminowej prognozy zapotrzebowania na energię elektryczną dla modeli neuronowych i neuronowo-rozmytych, pokazaliśmy, że w rozważanych przypadkach możemy przyjąć założenie o normalnym charakterze tego rozkładu (podrozdział 3.4). Takie założenie będziemy przyjmować również w bieżącym rozdziale. Przypomnieć trzeba jeszcze, że założenie to powinno zostać zweryfikowane w każdym przypadku odrębnie – poprzez sprawdzenie rozkładu reszt modelu.

Powtórzmy, że prognoza zapotrzebowania na energię $f(\mathbf{x}, \mathbf{w})$ wyznaczona dla tego problemu wydaje się pomyślna, zapotrzebowanie kształtuje się wyraźnie poniżej założonego progu decyzyjnego g. Wiedząc więc, że rozkład prawdopodobieństwa prognozy zapotrzebowania na energię ma charakter rozkładu normalnego o wartości oczekiwanej określonej przez prognozę oraz odchyleniu standardowym wyznaczonym metodami omawianymi w rozdziale 3, przyjmiemy, że jest to rozkład $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$, czyli dla przykładowych danych $N(189\,000, 10\,000)$. Oszacujmy dla tego rozkładu przedział prognozy z wysokim prawdopodobieństwem α.



Rysunek 4.2.1. Przedziały prognozy i bezpieczeństwo (stabilność) wyboru alternatywy decyzyjnej. Granica decyzyjna leży poza przedziałem prognozy i decyzja jest stabilna (a). Granica decyzyjna leży w obrębie przedziału i pojawia się niepewność (b)
Źródło: opracowanie własne

Gdyby granica decyzyjna *g* znalazła się poza znalezionym przedziałem (sytuacja zilustrowana na rysunku 4.2.1 (a)), oznaczałoby to, że leży ona w obszarze, w którym prawdopodobieństwo realizacji rzeczywistego zapotrzebowania na energię *ZE* jest już bardzo małe. W związku z tym oparcie naszej decyzji na założeniu, że zapotrzebowanie na energię kształtować się będzie poniżej *g*, należy uznać za poprawne i czynnik ryzyka może być pomijany.

Jeżeli natomiast granica decyzyjna *g* znajdzie się gdzieś wewnątrz przedziału (rysunek 4.2.1 (b)), to istnieje znaczące prawdopodobieństwo, że możemy ją jednak przekroczyć. Prawdopodobieństwo to jest tym większe, im bardziej wartość *g* oddala się od granicy przedziału prognozy i zbliża się do samej wartości prognozy $f(\mathbf{x}, \mathbf{w})$. W takim przypadku należy oczywiście z dużą nieufnością podchodzić do naszego warunku związanego z przewidywanym zapotrzebowaniem na energię i do oczekiwań odnośnie do sukcesu przy wyborze ryzykownego wariantu postępowania *AR*.

Przypomnijmy, że dla danego rozkładu prawdopodobieństwa prognozy $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$ przedział prognozy z prawdopodobieństwem α wyznaczyć możemy za pomocą kwantyli tego rozkładu *Q*:

$$[Q_{N(f,\sigma(\mathbf{x}))}((1-\alpha)/2), \quad Q_{N(f,\sigma(\mathbf{x}))}((1+\alpha)/2)] \tag{4.2.7}$$

Jeżeli dla przykładowych danych z problemu 4.2.2 wyznaczymy przedział prognozy zapotrzebowania na energię z prawdopodobieństwem α = 98%, korzystając z kwantyli rozkładu normalnego $N(189\,000, 10\,000)$, to otrzymamy przedział wartości zapotrzebowania w granicach [165 737, 212 263] kWh. Gdyby nasza granica decyzyjna *g* była większa i leżała poza górną granicą tego przedziału, to z 99-procentową pewnością moglibyśmy twierdzić, że zapotrzebowanie na energię będzie poniżej niej (poniżej i powyżej 98% przedziału prognozy pozostają obszary zapotrzebowania na energię obejmujące po 1% masy prawdopodobieństwa). Niestety tak nie jest. W rozważanej sytuacji prognoza zapotrzebowania na energię nie redukuje w pełni niepewności związanej z poziomem zapotrzebowania i ewentualnym sukcesem w przypadku wyboru rozwiązania ryzykownego *AR*. W związku z tym sama prognoza, w której została określona wyłącznie wartość oczekiwana zapotrzebowania, nie powinna stanowić jedynej podstawy decyzji. Powinniśmy zastosować metody wspomagania decyzji uwzględniające pojawiającą się niepewność oraz związany z nią element ryzyka.

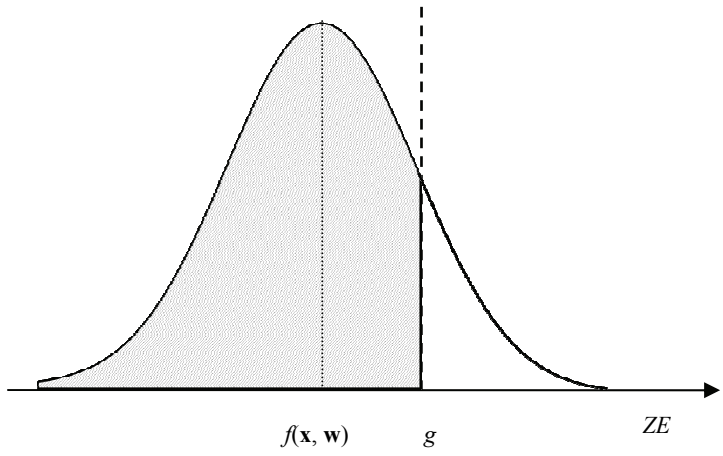
Aby podjąć właściwą decyzję, musimy oszacować czynnik ryzyka efektów jej podjęcia. W przypadku rozważanego problemu 4.2.1, analizując niepewność związaną ze skutkami wyboru rozwiązania ryzykownego *AR*, musimy określić, jakie jest prawdopodobieństwo *p* odniesienia sukcesu i $1 - p$ porażki, czyli zdarzenia polegającego na tym, że przy danej prognozie zapotrzebowanie na energię elektryczną ukształtuje się na poziomie nieprzekraczającym zakładanego

poziomu g ($ZE \le g$) lub powyżej niego ($ZE > g$). I znów – możemy to zrobić bez większych kłopotów, pod warunkiem jednak, że wykorzystamy nie tylko samą prognozę wartości oczekiwanej zapotrzebowania na energię $f(\mathbf{x}, \mathbf{w})$, dostarczaną przez sieć neuronową czy też neuronowo-rozmytą, ale za pomocą metod prezentowanych wcześniej w rozdziale 3 wyznaczymy odchylenie standardowe danej prognozy $\sigma(\mathbf{x})$, określając jej rozkład prawdopodobieństwa $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$.

Jeżeli spojrzymy na rysunek 4.2.2, to widzimy, że aby znaleźć prawdopodobieństwo p sukcesu ryzykownego rozwiązania AR, to jest zdarzenia $ZE \le g$, musimy znaleźć pole obszaru ograniczonego krzywą gęstości rozkładu prognozy, prostą $ZE = g$, oraz osią układu współrzędnych (obszar pokazany na rysunku jako zacieniowany). Otrzymujemy zatem:

$$p = \Pr(ZE \le g) = \int_{-\infty}^{g} p_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(y)dy = P_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(g) \tag{4.2.8}$$

gdzie przez $P_{N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))}$ oznaczyliśmy dystrybuantę rozkładu prognozy zapotrzebowania na energię $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$.



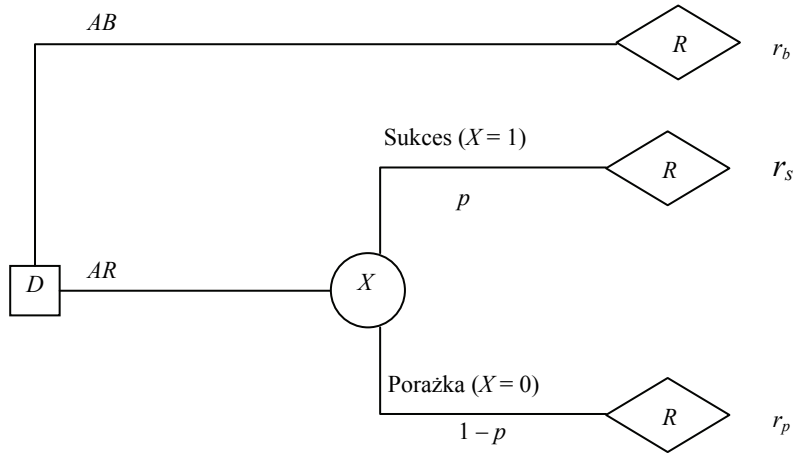
Rysunek 4.2.2. Ilustracja graficzna sposobu określania prawdopodobieństwa utrzymania się poniżej założonego progu g przy danym rozkładzie prognozy $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$
Źródło: opracowanie własne

Przypomnijmy, że w przykładowych danych do problemu 4.2.2 mamy określone odchylenie standardowe prognozy zapotrzebowania na energię $\sigma(\mathbf{x})$ = 10 000 kWh, a co za tym idzie rozkład rozważanej prognozy N(189 000, 10 000). Korzystając z tej informacji, bez trudu możemy obliczyć prawdopodo-

bieństwo p. Klasyczny sposób postępowania, zazwyczaj przedstawiany na kursach statystyki, polega na normalizacji wartości progu g:

$$g_{norm} = \frac{g - f(\mathbf{x}, \mathbf{w})}{\sigma(\mathbf{x})} = \frac{200000 - 189000}{10000} = 1{,}1 \tag{4.2.9}$$

a następnie odczytaniu wartości dystrybuanty rozkładu normalnego standardowego w punkcie $g_{norm} = 1{,}1$ z tabel statystycznych. Obecnie niemal każde oprogramowanie służące do obliczeń biznesowych, nawet tak proste jak arkusz kalkulacyjny, ma możliwość znajdowania wartości dystrybuanty dowolnego rozkładu normalnego. Bez trudu możemy więc obliczyć prawdopodobieństwo utrzymania się zapotrzebowania na energię *ZE* na poziomie nie wyższym niż $g = 200\,000$ kWh, $p \approx 86{,}4\%$. Oczywiście prawdopodobieństwo przekroczenia przez zapotrzebowanie na energię założonego poziomu wynosi $1 - p$, czyli około 13,6%.



Rysunek 4.2.3. Drzewo decyzyjne dla prostego wyboru między alternatywami bezpieczną *AB* i ryzykowną *AR*
Źródło: opracowanie własne

Zauważmy, że te informacje na temat prawdopodobieństw już same w sobie są dosyć istotne dla decydenta, który musi podjąć decyzję określoną w problemie 4.2.2. Nie otrzymuje on wyłącznie „suchej" wartości punktowej prognozy, ale dodatkowo widzi, jak rozkłada się ryzyko ewentualnego sukcesu i porażki przy wyborze ryzykownej alternatywy decyzyjnej *AR*. W naszym przypadku od

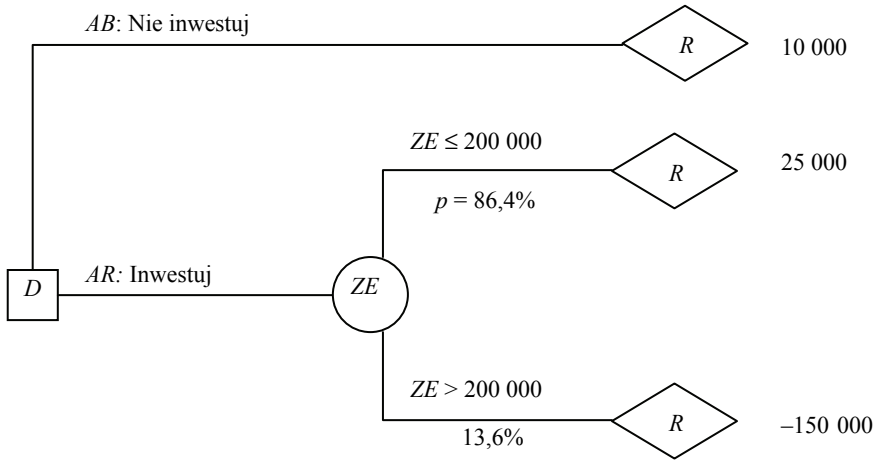
razu może więc zaobserwować, że prognoza zapotrzebowania na energię kształtuje się, co prawda, korzystnie – poniżej granicy decyzyjnej g, ale nie jest idealna. Prawdopodobieństwo, że zapotrzebowanie rzeczywiście utrzyma się poniżej tej granicy jest duże i wynosi około 86%. Istnieje jednak również mniej więcej 14-procentowe prawdopodobieństwo (ryzyko), że zapotrzebowanie na energię przekroczy jednak założony próg i wybór decyzji ryzykownej *AR* zakończy się porażką. Już choćby dzięki temu decydent otrzymuje pewne narzędzie do oceny, czy taki rozkład prawdopodobieństwa i ryzyko niepowodzenia są dla niego akceptowalne – z punktu widzenia jego indywidualnej skłonności do ryzyka – czy nie.

Obiektywny sposób porządkowania alternatyw w podobnych sytuacjach stanowi zasada wyboru decyzji przy wykorzystaniu reguły optymalizacji wartości oczekiwanej ich wyników.

Jeżeli niepewność wykorzystywanych danych powoduje, że wynik poszczególnych alternatyw decyzyjnych jest zmienną losową, to znając rozkład prawdopodobieństwa niepewności, możemy obliczyć jego wartość oczekiwaną. Jeśli skutki decyzji określone są w kategoriach zysków, to wybieramy rozwiązanie oferujące najwyższą wartość oczekiwaną zysku, jeżeli zaś w kategoriach kosztów, to przy wyborze kierujemy się regułą minimalizacji wartości oczekiwanej kosztu.

W przypadku ogólnym sytuację przedstawioną w problemie 4.2.2 możemy zobrazować za pomocą drzewa decyzyjnego znajdującego się na rysunku 4.2.3. Przyjmijmy, że mamy do wyboru dwie alternatywy decyzyjne: bezpieczną *AB*, której skutek r_b znany jest z góry, oraz ryzykowną *AR*, której skutek zależy od binarnej zmiennej losowej i wynosi r_s w przypadku sukcesu (z prawdopodobieństwem p) lub r_p w przypadku porażki (z prawdopodobieństwem $1 - p$). Zakładamy oczywiście, że $r_p < r_b < r_s$. Prawdopodobieństwo sukcesu równe jest p.

Drzewo decyzyjne składa się więc w tym przypadku z węzła decyzyjnego, z którego wychodzą gałęzie odpowiadające możliwym alternatywom decyzyjnym. Decyzja bezpieczna *AB* prowadzi bezpośrednio do węzła wypłaty r_b, ponieważ jej wybór niezależnie od okoliczności skutkuje takim zyskiem. Decyzja ryzykowna *AR* prowadzi do węzła zmiennej losowej X, od której zależy jej wynik. Gałęzie wychodzące z węzła X odpowiadają wszystkim możliwym stanom (wartościom) tej zmiennej losowej. I tak w przypadku sukcesu ($X = 1$) z prawdopodobieństwem wynoszącym p dochodzimy do węzła wyniku r_s, zaś w przypadku porażki – do wyniku r_s. Prawdopodobieństwo porażki ($X = 0$) równe będzie oczywiście $1 - p$.



Rysunek 4.2.4. Drzewo decyzyjne dla problemu 4.2.2
Źródło: opracowanie własne

Przypomnijmy, że w przypadku konkretnym dla omawianego zagadnienia podejmujemy decyzję o pewnej inwestycji, zaś rozważaną zmienną wpływającą na jej wynik jest wielkość zapotrzebowania na energię *ZE*, a konkretnie informacja, czy utrzyma się ono poniżej poziomu $g = 200\,000$ kWh (sukces), czy też go przekroczy (porażka). Za pomocą stworzonego modelu neuronowego (neuronowo-rozmytego) określiliśmy prognozowany rozkład zapotrzebowania jako rozkład normalny $N(189\,000, 10\,000)$ i przy jego użyciu wyznaczyliśmy prawdopodobieństwo $p = 86{,}4\%$. Wartości poszczególnych wyników wynoszą odpowiednio: $r_s = 25\,000$ złotych, $r_p = -150\,000$ złotych, $r_b = 10\,000$ złotych. Drzewo odpowiadające bezpośrednio decyzji określonej w problemie 4.2.2 zostało przedstawione na rysunku 4.2.4.

Spoglądając na rysunek 4.2.3 lub 4.2.4, widzimy od razu, że wartość oczekiwana wyniku bezpiecznej alternatywy decyzyjnej *AB* wynosi r_b, jako że jest to wartość stała, niezależna od okoliczności. Natomiast wartość oczekiwana wyniku alternatywy ryzykownej *AR* jest równa $pr_s + (1 - p)r_p$. Ponieważ oceniamy możliwe rozwiązania w kategoriach uzyskiwanego przy ich wyborze zysku, jako racjonalni decydenci wybierzemy oczywiście to, które zapewnia jego wyższą wartość. Ogólnie zatem reguła wyboru decyzji dla sytuacji opisanych za pomocą drzewa decyzyjnego na rysunku 4.2.3 ma postać:

$$\begin{aligned} &\text{wybieramy } AB, \text{ jeżeli: } r_b > pr_s + (1-p)r_p \\ &\text{wybieramy } AR, \text{ jeżeli: } r_b < pr_s + (1-p)r_p \end{aligned} \qquad (4.2.10)$$

Wybieramy więc tę alternatywę decyzyjną, dla której wartość oczekiwana uzyskiwanego wyniku jest wyższa. Gdyby w warunkach (4.2.10) występowała równość, wybór byłby obojętny.

Stosując więc warunki (4.2.10) do naszego przykładowego problemu 4.2.2, możemy wyznaczyć $E(ZE, AR)$, czyli wartość oczekiwaną skutku decyzji ryzykownej AR dotyczącej realizacji planowanej inwestycji przy danym prognozowanym rozkładzie zapotrzebowania na energię elektryczną ZE:

$$E(ZE, AR) = pr_s + (1 - p)r_p = \quad (4.2.11)$$
$$= 25\,000 \cdot 86{,}4\% + (-150\,000) \cdot 13{,}6\% \approx 1258$$

Porównajmy teraz uzyskany wynik z wartością oczekiwaną wyniku bezpiecznej alternatywy decyzyjnej AB o rezygnacji z inwestycji. Wynosi on $E(ZE, AB) = 10\,000$. Widzimy, że wyższą wartość oczekiwaną zysku oferuje rozwiązanie bezpieczne, więc to je powinniśmy wybrać, rezygnując z realizacji opcji ryzykownej.

Ten wniosek może wydawać się nieco zaskakujący w świetle naszych wcześniejszych rozważań. Wydawało się bowiem, że prognoza zapotrzebowania na energię jest dosyć korzystna, wyraźnie niższa od przyjętego limitu wpływającego na skutki analizowanej decyzji. Gdy ocenialiśmy prognozę w kategoriach błędów o charakterze czysto statystycznym, również wydawało się, że wybór alternatywy decyzyjnej jest raczej oczywisty i powinniśmy podjąć decyzję o zaangażowaniu się w rozważaną inwestycję. Omawiana sytuacja stanowi zatem modelowy przykład słuszności wcześniej sygnalizowanej już kilkukrotnie uwagi. Ocena prognozy w kategoriach czysto statystycznych, w oderwaniu od sytuacji decyzyjnej, w której będzie ona wykorzystywana, może okazać się niewystarczająca. Poprawne wykorzystanie prognozy wymaga rozważenia prawdopodobieństw skutków decyzji wynikających z niepewności przewidywanej zmiennej oraz skonfrontowania tych prawdopodobieństw z efektami, jakie przynosi wybór danego rozwiązania dla przyjętych przez decydenta celów. Stąd więc istotne jest, aby prognoza dostarczała narzędzi do takich analiz.

Aby wyjaśnić to do końca, powiedzmy wyraźnie, jaka jest interpretacja zasady podejmowania decyzji przy użyciu kryterium wartości oczekiwanej wyniku wybieranej alternatywy decyzyjnej. Zakładamy więc, że kierujemy się wartością oczekiwaną (4.2.11). Gdybyśmy wielokrotnie podejmowali takie same lub podobne (w podobnych warunkach) decyzje jak w problemie 4.2.2, to wybierając ryzykowną alternatywę decyzyjną AR o podjęciu inwestycji, znacznie częściej zarabialibyśmy 25 000 złotych, ponieważ prognoza jest dosyć pewna. Ale od czasu do czasu wybrane rozwiązanie skutkowałaby porażką, co dawałoby nam w wyniku dużo większą stratę, bo aż –150 000 złotych. Przeciętnie

z decyzji ryzykownej możemy spodziewać się zysku w wysokości 1 258 złotych, a więc wyraźnie niższego niż 10 000 złotych zysku otrzymywanego w przypadku wyboru alternatywy bezpiecznej.

Istotą reguły (4.2.10), podjęcia decyzji na podstawie wartości oczekiwanej jej skutków, jest więc wyważenie wszystkich możliwych efektów wyboru alternatywy decyzyjnej oraz prawdopodobieństw ich realizacji. Pozwala nam to ocenić niepewność prognozy w świetle uwarunkowań wynikających z zadania decyzyjnego, w którym jest ona wykorzystywana. Często wartość oczekiwaną oraz regułę wyboru rozwiązania za pomocą tej wartości określa się jako „wolną od nastawienia do ryzyka". Chodzi o to, że decydent kierujący się tą regułą zachowuje obiektywizm, nie kieruje się określonym nastawieniem psychologicznym – jest obojętny z punktu widzenia skłonności do ryzyka lub jej braku (optymizmu lub pesymizmu). Na zimno wyważa swoje szanse osiągnięcia określonych efektów oraz ich ewentualną skalę.

Reguła wartości oczekiwanej minimalizuje ryzyko niepożądanych skutków decyzji. Kryterium to stanowi naturalnie tylko pewną wskazówkę dla decydenta odnośnie do oceny alternatyw decyzyjnych. Analizując prawdopodobieństwa poszczególnych skutków decyzji, może on pokierować się inną strategią wyboru, bardziej pesymistyczną albo optymistyczną. Należy jednak zwrócić uwagę, że w takim przypadku decyzja ma charakter subiektywny, zależny od aktualnego stanu psychologicznego decydenta. Bardziej optymistyczne, agresywne strategie wyboru narażają nas na ryzyko porażki i wiążących się z nią gorszych wyników. Strategie bardziej zachowawcze, pesymistyczne zwiększają z kolei ryzyko niepełnego wykorzystania pojawiających się możliwości, jeśli dochodzi do sukcesu. Reguła wartości oczekiwanej, obiektywnie rzecz biorąc, ma charakter optymalny i nie jest obciążona tego rodzaju dodatkowym ryzykiem. Jeśli odchodzimy od niej, powinniśmy być zatem świadomi wiążącego się z tym ryzyka.

Niezależnie od zastosowanej reguły wyboru wprowadzanej w życie alternatywy decyzyjnej, widzimy, jak istotnym elementem i cennym narzędziem dla decydenta są prawdopodobieństwa realizacji różnych możliwych stanów rzeczywistości w świetle wykorzystywanej prognozy. Prognoza powinna więc dostarczać odpowiednich środków do poprawnej analizy skutków decyzji – takich, które pozwalają na szacowanie niepewności zależnych od niej faktów i zdarzeń.

Pojawia się wobec tego istotne pytanie, dla jakiego zakresu wartości prawdopodobieństwa sukcesu p optymalny będzie wybór bezpiecznej alternatywy decyzyjnej AB, a dla jakiego powinniśmy raczej wybrać alternatywę ryzykowną AR. Wyznaczone na podstawie prognozy prawdopodobieństwo utrzymania się zapotrzebowania na energię poniżej przyjętego progu g jest przecież tylko

oszacowaniem i chcielibyśmy wiedzieć, czy mamy jakiś margines prawdopodobieństwa dla naszej decyzji i czy z tego punktu widzenia jest ona „bezpieczna".

Graniczny moment, dla którego następuje zmiana wybieranej optymalnej alternatywy decyzyjnej, stanowi sytuacja, w której wartości oczekiwane wyników uzyskiwanych dla obu rozważanych wariantów decyzji, są równe. Otrzymujemy wówczas następujące równanie:

$$\begin{aligned} pr_s + (1-p)r_p &= r_b \\ pr_s + r_p - pr_p &= r_b \\ p(r_s - r_p) &= r_b - r_p \end{aligned} \tag{4.2.12}$$

Na podstawie (4.2.12) możemy więc wyznaczyć prawdopodobieństwo równowagi p^*, dla którego następuje powyższa zmiana decyzji. Wynosi ono oczywiście:

$$p^* = \frac{r_b - r_p}{r_s - r_p} \tag{4.2.13}$$

Jak widzimy, wartość granicznego prawdopodobieństwa równowagi p^* wynika bezpośrednio ze spodziewanych skutków decyzji. Jest to prosty stosunek tego, co możemy uzyskać, wybierając bezpieczną alternatywę decyzyjną, w porównaniu z najgorszą możliwością, do tego, co możemy uzyskać, wybierając rozwiązanie ryzykowne. Regułę (4.2.10) wyboru optymalnej alternatywy decyzyjnej, opartą na maksymalizacji wartości oczekiwanej wyniku, możemy więc sformułować także następująco: jeżeli oszacowane w danym przypadku prawdopodobieństwo sukcesu p jest większe od p^*, to powinniśmy podjąć decyzję ryzykowną, jeżeli zaś mniejsze – decyzję bezpieczną. W przypadku równości wybór jest, rzecz jasna, obojętny.

Dla naszego problemu 4.2.2 prawdopodobieństwo równowagi p^* możemy obliczyć ze wzoru (4.2.13) następująco:

$$p^* = \frac{r_b - r_p}{r_s - r_p} = \frac{10000 - (-150000)}{25000 - (-150000)} = \frac{260000}{275000} \approx 0{,}914 \tag{4.2.14}$$

Jak widzimy, w rozważanym przez nas problemie 4.2.2, prawdopodobieństwo równowagi wynosi około 91,4% i jest ono jednak nieco wyższe niż wynikające z naszej prognozy prawdopodobieństwo faktu, że zapotrzebowanie na energię ukształtuje się poniżej poziomu g = 200 000 kWh. Przypomnijmy, że wynosiło ono 86,4%.

Aby zakończyć temat prognoz zapotrzebowania na energię jako dyskretnych zmiennych losowych, przyjrzyjmy się jeszcze sytuacji, kiedy możliwe są więcej niż dwie wartości (stany rzeczywistości) tego rodzaju zmiennej. Podobnie jak poprzednio, przeanalizujemy to na przykładowym problemie praktycznym, który stanowi stosunkowo prostą modyfikację wcześniej rozważanych zagadnień.

Problem 4.2.3

Powróćmy jeszcze raz do decyzji opisanej w problemach 4.2.1 i 4.2.2. Przypuśćmy jednak, że po dokładniejszym rozpoznaniu sytuacji stwierdziliśmy, że przekroczenie progu g zapotrzebowania na energię ZE niekoniecznie musi od razu powodować tak wysokie straty w wysokości r_p. Istnieje pewien margines zapotrzebowania między g a pewnym poziomem g_1 ($g_1 > g$), w którym to obszarze jesteśmy sobie w stanie częściowo poradzić, ponosząc mniejsze straty (lub uzyskując niższe zyski niż w przypadku sukcesu) o wartości r_w. Dużą stratę r_p ponosimy dopiero przy bardzo wysokim poziomie zapotrzebowania na energię, przekraczającym g_1. Zakładamy ponadto, że wartość r_w jest jednak niższa niż wynik rozwiązania bezpiecznego AB, tzn. $r_p < r_w < r_b < r_s$. Pozostałe warunki podejmowanej decyzji oraz dane dotyczące prognozy zapotrzebowania na energię i jej rozkładu pozostają bez zmian.

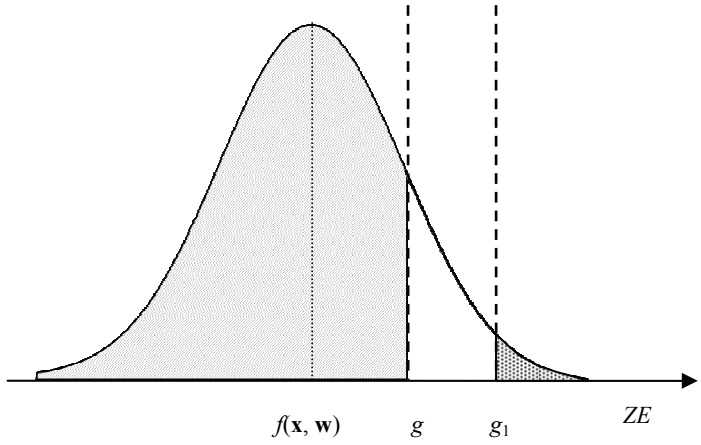
Przyjmijmy przykładowe wartości – nasz dodatkowy margines bezpieczeństwa zapotrzebowania na energię wynosi 10 000 kWh, czyli g_1 = 210 000 kWh oraz w tym zakresie ponosimy niewielką stratę w wysokości –5 000 złotych.

Ponownie korzystając z prognozowanego rozkładu zapotrzebowania na energię, musimy wyznaczyć prawdopodobieństwa poszczególnych możliwych stanów rzeczywistości wpływających na efekty rozważanej decyzji. Obecnie będziemy musieli, rzecz jasna, uwzględnić nie dwie, ale trzy możliwości: zapotrzebowanie na energię może ukształtować się na poziomie „normalnym", czyli gwarantującym sukces ($ZE \le g$), lub na poziomie „wysokim", ale jeszcze niepowodującym najgorszych strat ($g < \mathrm{ZE} \le g_1$), oraz na poziomie „bardzo wysokim" skutkującym porażką w najwyższym wymiarze.

Prawdopodobieństwo pierwszego z tych stanów, które oznaczmy przez p_s, jest takie samo jak prawdopodobieństwo sukcesu w poprzednio rozważanym przypadku i możemy je wyznaczyć jako $P_{N(f(\mathbf{x},\mathbf{w}),\,\sigma(\mathbf{x}))}(g)$, czyli wartość dystrybuanty prognozowanego rozkładu zapotrzebowania na energię $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$ dla pierwszego z poziomów granicznych g. W przypadku danych wykorzystanych w naszym przykładzie, to jest rozkładu prognozy $N(189\,000,\ 10\,000)$ oraz granicznego poziomu zapotrzebowania g = 200 000 kWh, prawdopodobieństwo to wynosi:

$$p_s = P_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(g) = P_{N(189000,10000)}(2000000) \approx 86{,}4\% \qquad (4.2.15)$$

Prawdopodobieństwo wysokiego (czyli pomiędzy poziomami g i g_1) poziomu zapotrzebowania energii, które będziemy oznaczać dalej przez p_w, odpowiada białemu polu pod funkcją gęstości na rysunku 4.2.5. Łatwo więc zauważyć, że możemy je wyznaczyć jako różnicę wartości dystrybuanty $P_{N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))}$ prognozowanego zapotrzebowania na energię $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$, dla obu rozważanych poziomów granicznych g_1 i g:



Rysunek 4.2.5. Prawdopodobieństwo przedziału $[g, g_1]$ wartości zmiennej przy danym rozkładzie prognozy $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$ stanowi biały obszar pod krzywą gęstości
Źródło: opracowanie własne

$$\begin{aligned} p_w &= \Pr(g < ZE \le g_1) = \int_{g}^{g_1} p_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(y)dy = \\ &= \int_{-\infty}^{g_1} p_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(y)dy - \int_{-\infty}^{g} p_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(y)dy = \\ &= P_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(g_1) - P_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(g) \end{aligned} \tag{4.2.16}$$

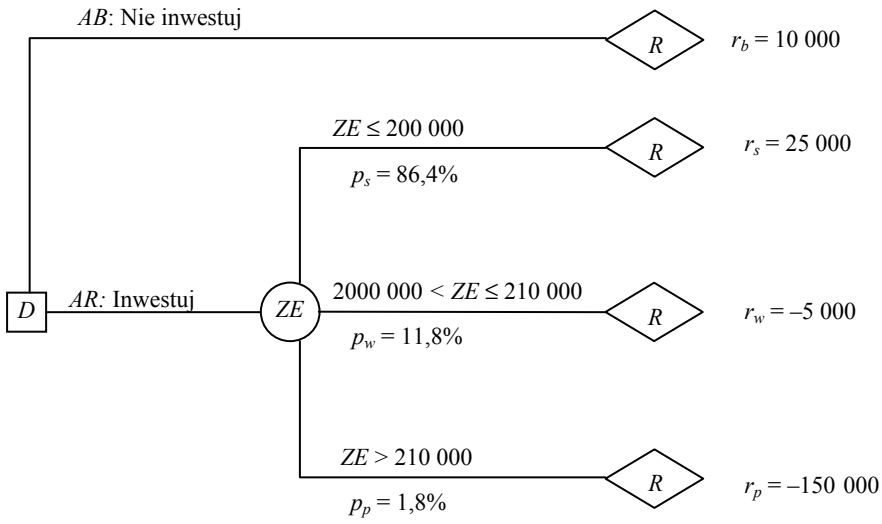
Dla rozkładu prognozy $N(189\,000,\ 10\,000)$ oraz granicznych poziomów zapotrzebowania na energię $g = 200\,000$ kWh i $g_1 = 210\,000$ kWh prawdopodobieństwo p_w wynosi więc:

$$p_w = P_{N(189000,10000)}(210000) - P_{N(189000,10000)}(200000) \approx 11{,}8\% \tag{4.2.17}$$

Ostatnie prawdopodobieństwo na poziomie „bardzo wysokim", przekraczającym wartość graniczną g_1 i skutkującym największą stratą r_p, możemy obliczyć już łatwo. Będzie ono równe po prostu dopełnieniu prawdopodobieństw

pozostałych przypadków do jedności. Oznaczmy więc powyższe prawdopodobieństwo przez p_p. W naszym przykładzie wynosi ono:

$$p_p \approx 1 - 86{,}4\% - 11{,}8\% = 1{,}8\% \tag{4.2.18}$$



Rysunek 4.2.6. Drzewo decyzyjne dla prostego zagadnienia decyzyjnego przedstawionego w problemie 4.2.3
Źródło: opracowanie własne

Jak więc widzimy, w zagadnieniu przedstawionym w problemie 4.2.3 prawdopodobieństwo poniesienia bardzo dużej straty jest znacznie mniejsze niż w przypadku decyzji omawianej w poprzednim problemie. Wybierając ryzykowną alternatywę decyzyjną *AR*, decydent musi obecnie liczyć się z następującymi możliwymi skutkami oraz prawdopodobieństwami ich realizacji: przy normalnym zapotrzebowaniu na energię r_s = 25 000 złotych, z prawdopodobieństwem p_s = 86,4%; przy wysokim r_w = –5 000 złotych, z prawdopodobieństwem p_w = 11,8% oraz przy bardzo wysokim r_w = –150 000 złotych, z prawdopodobieństwem p_p = 1,8%.

Proces analizy wynikających z tych warunków decyzji możemy przedstawić za pomocą drzewa decyzyjnego znajdującego się na rysunku 4.2.6. Jak widzimy, z węzła zmiennej losowej (zdarzenia losowego) *ZE*, wpływającej na skutki decyzji ryzykownej *AR*, wychodzą tym razem trzy gałęzie, które odpowiadają poszczególnym ich wartościom. Biorąc więc pod uwagę możliwe efekty decyzji ryzykownej oraz prawdopodobieństwa ich wystąpienia dla różnych możliwych stanów zapotrzebowania na energię elektryczną, wartości oczekiwane skutków

wariantów bezpiecznego $E(ZE, AB)$ oraz ryzykownego $E(ZE, AR)$ rozważanej w przykładowym problemie 4.2.3 decyzji możemy wyznaczyć:

$$\begin{aligned} E(ZE, AB) &= 10\,000 \\ E(ZE, AR) &= p_s r_s + p_w r_w + p_p r_p = \\ &= 25\,000 \cdot 86{,}4\% + (-5\,000) \cdot 11{,}8\% + \\ &+ (-150\,000) \cdot 1{,}8\% \approx 18\,340 \end{aligned} \tag{4.2.19}$$

Stosując jako kryterium wyboru optymalizację wartości oczekiwanej wyników decyzji, tym razem wybralibyśmy alternatywę ryzykowną *AR*. Jak widzimy, oczekiwane skutki wyboru tego wariantu są lepsze niż w przypadku rozwiązania bezpiecznego. W porównaniu z decyzją rozważaną w problemie 4.2.2 poziom zapotrzebowania na energię rzadziej będzie na tyle wysoki, aby spowodować stratę dużej kwoty r_p. Bardziej więc opłaca nam się zaryzykować, a wartość oczekiwana zysku dla decyzji o podjęciu inwestycji jest wyższa niż dla decyzji o rezygnacji z niej.

Decydent naturalnie wciąż może odejść od reguły wyboru alternatywy decyzyjnej na podstawie wartości oczekiwanej wyniku, ale musi być świadom, że postępowanie takie ma charakter subiektywny, związany z jego nastawieniem psychologicznym, indywidualną skłonnością do ryzyka i optymistycznym lub pesymistycznym postrzeganiem sytuacji. Odejście od obiektywnego kryterium wartości oczekiwanej, jak już podkreślaliśmy wcześniej, naraża na dodatkowe ryzyko związane z jednej strony z dużą porażką, a z drugiej strony – z ewentualnymi niewykorzystanymi możliwościami. Nawet wówczas prawdopodobieństwa realizacji poszczególnych możliwych stanów rzeczywistości (a co za tym idzie skutków wyboru) stanowią cenną informację pozwalającą na lepszą orientację decydenta w czynnikach ryzyka związanego z podejmowaną decyzją.

Jak więc widzimy, w praktycznych zagadnieniach decyzyjnych często może zachodzić potrzeba przekształcenia ciągłych wartości prognozowanych zmiennych w dyskretne zdarzenia losowe, które wpływają na wyniki podejmowanej decyzji. Istotną rolę w takim przypadku odgrywa możliwość rozważenia wykorzystania, jako źródła informacji, danych o niepewności prognozy. Pozwala to na określenie stabilności wyboru, a w warunkach ryzyka – na określenie prawdopodobieństw poszczególnych stanów rzeczywistości wynikających z prognozy.

Do wskazanych celów nie wystarcza wykorzystywanie wyłącznie predykcji punktowej w formie wartości oczekiwanej zapotrzebowania na energię. Niezbędne jest wzięcie pod uwagę niepewności prognozy w postaci jej rozkładu prawdopodobieństwa. Dlatego w przypadku wykorzystania modeli prognostycznych w formie regresorów, takich jak sieci neuronowe czy neuronowo-rozmyte, metody szacowania rozkładu prognozowanego zapotrzebowania na energię, omawiane i analizowane w rozdziale 3 naszej pracy, mają kluczowe znaczenie

– pozwalają one na odpowiednią analizę skutków podejmowanej decyzji, elementów ryzyka, jakie się z nią wiąże, oraz na właściwy wybór najlepszego sposobu postępowania w danej sytuacji.

4.2.2. Prognozy zapotrzebowania na energię jako ciągłe zmienne losowe

W poprzednim punkcie analizowaliśmy zagadnienia decyzyjne, w których skutki podejmowanych decyzji zależne były w sposób skokowy od różnych poziomów zapotrzebowania na energię, jakie mogły zrealizować się w przyszłości. Nasze zyski, straty, koszty czy inne cele (kryteria), ze względu na które oceniamy efekty wybieranych przez decydenta alternatyw decyzyjnych, zmieniały się w sposób skokowy i miały w określonych przedziałach zapotrzebowania stałą wartość.

Obecnie przyjrzymy się nieco odmiennej sytuacji, w której efekty i skutki podejmowanej decyzji zmieniają się w sposób ciągły, w zależności od rzeczywistej wartości zapotrzebowania na energię. Natomiast, co ważne, nadal zakładać będziemy skończoną liczbę rozważanych alternatyw decyzyjnych. Innymi słowy, chwilowo nie interesują nas zagadnienia wyznaczania optymalnej wielkości rozmaitych ciągłych zmiennych decyzyjnych opartych na prognozie zapotrzebowania (jak np. wielkość zamówienia). Tego typu problemy będą analizowane w podrozdziale 4.3. Dokonujemy więc wyboru spośród skończonej liczby alternatyw decyzyjnych, których skutki (wyniki) opisane są ciągłymi rozkładami prawdopodobieństwa zależnymi od prognozowanego zapotrzebowania na energię.

I znów oczywiście nie należy traktować naszych rozważań w bieżącym punkcie jako obszernego i kompletnego wykładu z zakresu powyższych zagadnień. Podobnie jak i w punkcie poprzednim, interesuje nas raczej przedstawienie podstawowych kwestii związanych z powiązaniem regresorów, takich jak sieci neuronowe czy neuronowo-rozmyte, z zadaniami decyzyjnymi oraz wykazanie istotnego znaczenia analizowanych w punkcie 3 metod wyznaczania rozkładu prawdopodobieństwa zapotrzebowania na energię dla tego rodzaju modeli w praktycznych zagadnieniach na rynku energii.

Ponieważ nadal mówimy o decyzjach w warunkach ryzyka, podstawową regułą wyboru najlepszej alternatywy decyzyjnej pozostaje optymalizacja wartości oczekiwanej skutków decyzji w zależności od zmieniającego się zapotrzebowania na energię. Nie możemy jednak, tak jak w poprzednim punkcie, przyjąć żadnego stałego poziomu zysku czy kosztów wyznaczonych dla różnych przedziałów zapotrzebowania. Skutki rozważanych wariantów danej decyzji są ciągłymi zmiennymi losowymi, musimy więc rozważyć ich niepewność w postaci ciągłego rozkładu prawdopodobieństwa, generowanego przez

rozkład prognozy obciążenia sieci, oraz wyznaczać wartość kryterium każdej alternatywy decyzyjnej w formie wartości oczekiwanej rozkładu ciągłego.

Podobnie jak w poprzednim punkcie, to zagadnienie przeanalizujemy na przykładowym problemie decyzyjnym. Przedstawiony przykład ma naturalnie również charakter poglądowy i posłuży nam do zilustrowania sytuacji, w której niezbędne jest wzięcie pod uwagę niepewności prognozy zapotrzebowania na energię oraz przedstawienie sposobu wykorzystania modelu tej niepewności w ocenie ryzyka efektów rozważanych alternatyw decyzyjnych. Zastanówmy się więc nad następującym zagadnieniem:

Problem 4.2.4

Przeanalizujmy sytuację, w której przedsiębiorstwo zajmujące się obrotem energią elektryczną rozważa dwa różne warianty zakupu energii różniące się od siebie warunkami płatności za tę transakcję:

– w wariancie pierwszym płatność za odebraną energię odbywa się po jednolitej cenie zakupu $c_1 = 195$ złotych za megawatogodzinę (MWh),

– w wariancie drugim możemy kupić energię o 10 złotych taniej, w cenie $c_{21} = 185$ złotych za MWh; warunki te dotyczą jednakże jedynie pakietu energii nieprzekraczającego pewnego poziomu granicznego $g = 500$ MWh; jeżeli zapotrzebowanie ZE przekroczy poziom g, za dodatkową energię niezbędną do jego pokrycia, musimy zapłacić znacznie wyższą cenę $c_{22} = 400$ złotych za MWh.

Widzimy więc, że schemat sytuacji decyzyjnej jest dosyć zbliżony do zagadnienia rozważanego w problemie 4.2.1, z tą różnicą, że w obecnym przypadku decyzja nie skutkuje stałymi wynikami, zależnymi od przekroczenia (lub nie) pewnego poziomu zapotrzebowania na energię g. Zmieniają się one dla każdej rzeczywistej wartości zapotrzebowania ZE, jaka zrealizuje się w przyszłości. Dla obydwu rozwiązań możemy zdefiniować funkcje kosztów (ponieważ w tym przypadku kryterium oceny są koszty zakupu energii) $k_1(y)$ i $k_2(y)$. Przez y oznaczamy konkretną wartość zapotrzebowania na energię (jej zakupu) ZE.

Dla wariantu pierwszego funkcja kosztów zakupu energii jest prostą funkcją liniową:

$$k_1(y) = c_1 y = 195y \tag{4.2.20}$$

Dla drugiego wariantu funkcja kosztów zakupu jest nieco bardziej złożona, musimy bowiem rozważyć dwie różne ceny: c_{21} – gdy zapotrzebowanie y jest mniejsze lub równe od g, oraz c_{22} – jeżeli występuje nadwyżka ponad g. Funkcję tę możemy zdefiniować w następujący sposób:

$$k_2(y) = c_{21}\min(g, y) + c_{22}\max(y - g, 0) = \\ = 185\min(500, y) + 400\max(y - 500, 0). \tag{4.2.21}$$

I znów, podobnie jak w przypadku problemu 4.2.1, zastanówmy się najpierw nad sytuacją, w której decydent nie dysponuje żadną prognozą zapotrzebowania na energię. Ponieważ nie wie on, na jakim poziomie owo zapotrzebowanie się ukształtuje, podejmuje decyzję w warunkach niepewności. Jedynym sposobem analizy tej decyzji jest w zasadzie podejście symulacyjne.

Tabela 4.2.3. Tabela symulacji wyników (kosztów zakupu energii) dla poszczególnych alternatyw decyzyjnych w przypadku problemu 4.2.4 (w MWh)

Zapotrzebowanie na energię ($ZE = y$)	Wariant 1 $k_1(y)$	Wariant 2 $k_2(y)$
450	87 750	83 250
475	92 625	87 875
500	97 500	92 500
525	102 375	102 500
550	107 250	112 500
575	112 125	122 500

Źródło: opracowanie własne.

Ponieważ nie wiadomo, jaki stan rzeczywistości (poziom zapotrzebowania na energię) zrealizuje się w przyszłości, należy przeprowadzić symulację wyników poszczególnych alternatyw decyzyjnych dla różnych (oczywiście jak najbardziej reprezentatywnych) stanów oraz przeanalizować uzyskane wyniki za pomocą kryteriów wyboru w warunkach niepewności, omawianych w problemie 4.2.1.

Przyjmijmy, że decydent uważa, iż zapotrzebowanie na energię powinno ukształtować się w okolicy progu występującego w wariancie drugim zakupu, to jest 500 MWh, natomiast nie jest w stanie powiedzieć, jakie wartości w otoczeniu tej wielkości może ono przyjąć. W związku z tym przeanalizowane zostaną koszty obu wariantów zakupu dla wybranych reprezentatywnych stanów z tego otoczenia. W tabeli 4.2.3 znajdują się obliczone wartości funkcji kosztów $k_1(y)$ określonej przez (4.2.20) i $k_2(y)$ określonej przez (4.2.21).

Wykorzystanie zachowawczej i pesymistycznej reguły Walda polegałoby na wyborze alternatywy decyzyjnej, które daje najlepsze wyniki w przypadku realizacji najgorszego możliwego stanu rzeczywistości. Wyznaczamy więc dla każdego wariantu (kolumny w tabeli 4.2.3) wartość maksymalną jej kosztów. W naszym przypadku wynosi ona 112 125 złotych dla wariantu pierwszego

zakupu oraz 122 500 złotych dla wariantu drugiego. Kierując się regułą Walda, decydent pesymistyczny, myślący przede wszystkim o ograniczaniu ewentualnych strat, powinien więc wybrać wariant pierwszy zakupu energii, ponieważ w najgorszym przypadku skutkuje on, pomimo wszystko, niższymi kosztami. Zauważmy przy tym, że jeżeli uznajemy za możliwe ukształtowanie się stanu zapotrzebowania na energię na poziomie powyżej około 525 MWh, to kierując się regułą pesymistyczną, zawsze wybierzemy wariant pierwszy zakupu. Z danych znajdujących się w tabeli 4.2.3 wynika od razu, że dla coraz wyższej wartości zapotrzebowania y różnica w kosztach również rośnie.

Reguła optymistyczna polega na szukaniu przez decydenta przede wszystkim szans, które chce wykorzystać, nie oglądając się na ryzyko ewentualnego niepowodzenia. Polega więc ona na wyborze sposobu postępowania dającego najlepsze skutki przy realizacji najlepszego z możliwych stanów rzeczywistości. W rozważanym przypadku polegałoby to na znalezieniu w tabeli 4.2.3 wartości minimalnej kosztów dla każdego wariantu decyzji. Bez trudu możemy zauważyć, że wynosi ona 87 750 złotych dla wariantu pierwszego zakupu oraz 83 250 złotych dla wariantu drugiego. Kierując się tą regułą, decydent optymistyczny, skłonny do podjęcia ryzyka, powinien więc dla odmiany wybrać wariant drugi rozważanej decyzji, ponieważ przy tym najlepszym możliwym rozwoju wydarzeń wiąże się on z niższymi kosztami zakupu.

Zauważmy, że tym razem, jeżeli dopuszczamy możliwość ukształtowania się zapotrzebowania na energię na poziomie poniżej około 525 MWh, kierując się regułą optymistyczną, zawsze wybierzemy wariant drugi zakupu. Analizując dane w tabeli 4.2.3, widzimy tym razem, że dla coraz niższej wartości zapotrzebowania y różnica w kosztach między obydwoma wariantami kształtować się będzie zawsze na korzyść drugiego wariantu.

Przypomnijmy, że strategią pośrednią pozwalającą na wypracowanie kompromisu pomiędzy ekstremalnie przeciwstawnymi regułami pesymistyczną i optymistyczną jest metoda Hurwicza. Pozwala ona na pewne sterowanie czynnikiem skłonności do ryzyka, wykorzystując jako podstawę wyboru średnią ważoną z najlepszego i najgorszego wyniku każdego z rozważanych rozwiązań. Waga λ jest stałą z przedziału [0, 1] określającą zakładany poziom optymizmu (skłonności do ryzyka) decydenta. W przypadku kryterium problemu 4.2.4 oceny poszczególnych alternatyw decyzyjnych $kr(i)$, $i = 1, 2$ możemy więc zapisać jako:

$$kr(i) = \lambda \min_j k_i(y_j) + (1-\lambda) \max_j k_i(y_j) \tag{4.2.22}$$

Wybierając współczynnik skłonności do ryzyka odpowiednio do nastawienia decydenta, np. w przypadku decydenta dosyć optymistycznego i skłonnego do podjęcia ryzyka, dla $\lambda = 0{,}8$, otrzymujemy:

– dla wariantu pierwszego: $kr(1) = 0{,}8 \cdot 87\,750 + 0{,}2 \cdot 112\,125 = 92\,625$,

– dla wariantu drugiego: $kr(2) = 0{,}8 \cdot 83\,250 + 0{,}2 \cdot 122\,500 = 91\,100$.

W rozważanym przypadku decydent dosyć optymistyczny, jak widzimy, powinien wybrać drugi wariant zakupu energii.

Należy jednak zwrócić uwagę, że reguła Hurwicza daje w tym przypadku wyłącznie bardzo orientacyjne wyniki. Na kryterium (4.2.22) silny wpływ ma bowiem nie tylko fakt, która alternatywa decyzyjna daje relatywnie lepsze wyniki w najlepszej lub najgorszej sytuacji, ale także dokładne wartości tego minimum i maksimum. Ponieważ nie wiemy, czy tabela 4.2.3 obejmuje zakres wszystkich możliwych stanów zapotrzebowania, w związku z tym nie potrafimy dokładnie określić skutków decyzji w przeciwstawnych sytuacjach. Jeżeli dla przykładu możliwe są jeszcze wyższe stany zapotrzebowania na energię, to różnica w najgorszym (maksymalnym) wyniku obu wariantów będzie szybko rosnąć. Może się wówczas okazać, że przy tym samym współczynniku optymizmu (skłonności do ryzyka) $\lambda = 0{,}8$ nasza preferencja co do wyboru wariantu zakupu energii będzie zupełnie inna.

Przeanalizujmy jeszcze krótko, dla sytuacji decyzyjnej rozważanej w problemie 4.2.4, ostatnią z reguł wyboru alternatywy decyzyjnej w warunkach niepewności omawianych w punkcie 4.2.1. Przypomnijmy, że reguła, o której mówimy, Niehansa–Savage'a lub inaczej „reguła utraconej szansy", opiera się na minimalizacji możliwej straty albo żalu z tego powodu, że nie wybraliśmy najlepszego wariantu dla każdego stanu rzeczywistości.

Dla naszego przykładowego problemu 4.2.4 straty z utraconych szans wyznaczone dla stanów rzeczywistości oraz odpowiadających im kosztów obu wariantów zakupu energii z tabeli 4.2.3 znajdują się w tabeli 4.2.4. Analizując zawarte w tej tabeli dane, od razu w zasadzie możemy powiedzieć, że:

– dla wariantu pierwszego największa utracona szansa wynosi 5 000 złotych,

– dla wariantu drugiego największa utracona szansa wynosi 10 375.

Tabela 4.2.4. Tabela wielkości utraconych szans dla poszczególnych wariantów w przypadku problemu 4.2.4 (w MWh)

Zapotrzebowanie na energię ($ZE = y$)	Wariant 1 $k_1(y)$	Wariant 2 $k_2(y)$
450	4 500	0
475	4 750	0
500	5 000	0
525	0	125
550	0	5 250
575	0	10 375

Źródło: opracowanie własne.

Jak już pokazywaliśmy w poprzednim punkcie, „reguła utraconej szansy" Niehansa–Savage'a ma dosyć zachowawczy charakter, kierując się więc tym kryterium wybralibyśmy również wariant pierwszy zakupu energii. Ponadto widzimy, że jeżeli dopuszczamy jako możliwe stany zapotrzebowania przekraczające około 525 MWh, to reguła ta zawsze będzie prowadzić do preferowania wariantu pierwszego. W innym przypadku bardziej opłacalny będzie wariant drugi.

Ponownie przeanalizowaliśmy więc kilka reguł porządkowania i wyboru alternatywy decyzyjnej w warunkach niepewności. Z powodu braku informacji na temat istotnego zjawiska, jakim jest kształtowanie się poziomu zapotrzebowania na energię elektryczną, wykorzystujemy tutaj techniki symulacyjne, wyznaczając efekty możliwych rozwiązań dla różnych wartości realizacji popytu oraz analizując otrzymane wyniki pod kątem nastawienia psychologicznego decydenta, jego optymizmu i skłonności do ryzyka. Do dokładniejszego rozważenia istniejącej sytuacji decyzyjnej, przeanalizowania czynnika ryzyka, jaki się z nią wiąże, niezbędne jest jednak wykorzystanie prognoz lub oszacowań nieznanych informacji wpływających na efekty decyzji.

Prognoza zapotrzebowania na energię pozwala nam na redukcję niepewności związanej z tą zmienną poprzez narzucenie na zbiór możliwych wartości popytu ograniczenia w postaci jego przewidywanego rozkładu prawdopodobieństwa. Dzięki temu uzyskujemy możliwość bardziej precyzyjnego określenia skutków podejmowanej decyzji oraz przeanalizowania ryzyka rozważanych alternatyw decyzyjnych. Ponownie przyjrzyjmy się temu zagadnieniu na podstawie sytuacji, która stanowi rozszerzenie przykładowego problemu 4.2.4.

Problem 4.2.5

Przyjmijmy więc, że analizując decyzję dotyczącą wyboru wariantu zakupu energii opisaną w problemie 4.2.4, decydent chciał zredukować swoją niepewność odnośnie do wielkości zapotrzebowania na energię elektryczną. Wykorzystał w tym celu model prognostyczny oparty na sieci neuronowej lub neuronowo-rozmytej, uzyskując prognozę zapotrzebowania w rozważanym okresie o wartości $f(\mathbf{x}, \mathbf{w}) = 515$ MWh, gdzie przez $\mathbf{x}$ rozumiemy wzorzec danych wejściowych prognozy, natomiast przez $\mathbf{w}$ zestaw wag sieci. Odchylenie standardowe dla tej konkretnej prognozy, uzyskane jedną z metod analizowanych w rozdziale 3, zostało oszacowane na $\sigma(\mathbf{x}) = 85$ MWh.

Warunki płatności pozostają takie same jak w problemie 4.2.4, więc koszty w wariantach pierwszym i drugim zakupu opisywane są nadal przez funkcje kosztów $k_1(y)$ i $k_2(y)$, określone odpowiednio zależnością (4.2.20) oraz (4.2.21).

Przedstawione zagadnienie stanowi oczywiście przykład dużo ogólniejszej sytuacji decyzyjnej, w której skutki wyboru określone są za pomocą funkcji celu, zależnych od pewnej ciągłej zmiennej losowej. Ponownie analizując typowe sposoby wykorzystania przez decydentów posiadanych prognoz tej

zmiennej, należy powiedzieć, że często stosowanym podejściem w przypadku zbliżonym do problemu 4.2.5 jest po prostu zignorowanie niepewności prognozy. Wybór optymalnego sposobu postępowania polega wyłącznie na określeniu kryteriów poszczególnych wariantów decyzji poprzez podstawienie prognozy do odpowiednich wykorzystywanych przez decydenta funkcji celu (kosztów, zysków itp.). Konkretnie mówiąc, w przypadku naszego przykładowego problemu 4.2.5 ten schemat postępowania polega na podstawieniu otrzymanej prognozy zapotrzebowania na energię elektryczną do funkcji kosztów $k_1(y)$ i $k_2(y)$.

Dla wariantu pierwszego zakupu energii koszt wyznaczamy jako wartość funkcji kosztów $k_1(y)$ dla otrzymanej prognozowanej wartości zapotrzebowania na energię $y = f(\mathbf{x}, \mathbf{w}) = 515$ MWh:

$$k_1(f(\mathbf{x},\mathbf{w})) = k_1(515) = 195 \cdot 515 = 100425 \tag{4.2.23}$$

W przypadku drugiego wariantu zakupu energii wyznaczamy dla naszej prognozy wartość funkcji kosztów $k_2(y)$:

$$\begin{aligned} k_2(f(\mathbf{x},\mathbf{w})) = k_2(515) &= 185\min(500,515) + 400\max(515-500,0) = \\ &= 185 \cdot 500 + 400 \cdot 15 = 98500 \end{aligned} \tag{4.2.24}$$

Porównując obliczone dla prognozowanego zapotrzebowania na energię koszty obu wariantów rozważanej decyzji, wyznaczone przez zależności (4.2.23) oraz (4.2.24), wydaje się, że wybór jest dosyć jednoznaczny. Koszty wariantu drugiego zakupu energii elektrycznej są jednak nieco niższe niż w przypadku wariantu pierwszego. W świetle informacji dostarczonej przez prognozę zapotrzebowania na energię mogłoby się więc wydawać, że powinniśmy preferować właśnie tę drugą alternatywę decyzyjną.

Podobny sposób postępowania, jak już wcześniej nadmieniliśmy, oznacza zupełne zignorowanie faktu niepewności prognozy. W rozważanym przypadku, obliczając przy użyciu zależności (4.2.23) oraz (4.2.24) naszą funkcję celu, za pomocą której oceniamy brane pod uwagę alternatywy decyzyjne, czyli koszty zakupu energii, wykorzystujemy wyłącznie wartość oczekiwaną przyszłego zapotrzebowania. To prawda, że wynik prognoz punktowych zapotrzebowania, otrzymywanych z wykorzystaniem rozważanych w naszej pracy modeli regresyjnych, takich jak sieci neuronowe czy neuronowo-rozmyte, czyli wartość oczekiwana prognozowanej zmiennej określa najbardziej prawdopodobny poziom popytu na energię, ale to wcale przecież nie znaczy, że rzeczywista realizacja wartości zapotrzebowania będzie dokładnie odpowiadać otrzymanej prognozie.

Jeszcze raz bowiem zwróćmy tutaj uwagę na fakt, że prognoza nie stanowi informacji absolutnie pewnej i bezdyskusyjnej. Każda prognoza obarczona jest błędem i nie możemy oczekiwać pełnej zgodności rzeczywistej realizacji prognozowanego zjawiska z prognozą. Oczywiście w wielu przypadkach niepewność prognozy nie ma większego znaczenia dla podejmowanej decyzji. Zagadnienie to będziemy analizować zresztą dokładniej w dalszej części bieżącego punktu. Tym niemniej jednak bardzo często niepewność prognozy ma znaczenie wręcz podstawowe, w poważnym stopniu zmieniając ocenę rozważanych alternatyw decyzyjnych oraz wybór najlepszej z nich, najbardziej dostosowanej do konkretnej sytuacji, z którą mamy do czynienia.

Niepewność prognozy pociąga za sobą niepewność przyjętych funkcji celu, które służą do oceny skutków rozważanych alternatyw decyzyjnych. W naszym przypadku więc, aby uwzględnić niepewność prognozowanego poziomu zapotrzebowania na energię, musimy rozważyć wynikającą z niej niepewność kosztów zakupu energii. W tym celu właściwą metodą jest, podobnie jak w przypadku zagadnień rozważanych w poprzednim punkcie 4.2.1, zastosowanie kryterium wyboru alternatywy decyzyjnej, opartego nie bezpośrednio na funkcji celu, lecz na jej wartości oczekiwanej. Przypomnijmy, że wartość oczekiwana funkcji celu umożliwia wyważenie wszystkich możliwych wyników danego rozwiązania oraz prawdopodobieństw ich realizacji, stanowiąc obiektywną, wolną od jakiegoś konkretnego nastawienia do ryzyka, ocenę tego, co najbardziej prawdopodobnie możemy osiągnąć, podejmując daną decyzję.

W przypadku przykładowego problemu 4.2.5, aby uwzględnić niepewność prognozowanego poziomu popytu na energię elektryczną, wyboru wariantu zakupu energii powinniśmy więc dokonać na podstawie nie bezpośrednio funkcji kosztów $k_1(y)$ i $k_2(y)$, lecz wartości oczekiwanych tych kosztów dla danego rozkładu prawdopodobieństwa prognozy zapotrzebowania.

Musimy, rzecz jasna, wziąć pod uwagę jeden fakt. W poprzednim rozdziale efekty rozważanych alternatyw decyzyjnych były skokowymi zmiennymi losowymi. Wyznaczając zatem ich wartości oczekiwane, stosowaliśmy metody odpowiednie dla dyskretnych rozkładów prawdopodobieństwa. Obecnie mamy do czynienia z nieskończenie dużą liczbą możliwych stanów zapotrzebowania (a co za tym idzie kosztów każdego rozwiązania). Do wyznaczenia wartości oczekiwanej tych kosztów musimy więc zastosować metody charakterystyczne dla ciągłych zmiennych losowych.

Jeżeli przez $E(Y, R_A)$ oznaczymy wartość oczekiwaną wyniku R_A pewnej rozważanej alternatywy decyzyjnej A, obarczonego niepewnością wywoływaną przez rozkład prawdopodobieństwa pewnej zmiennej losowej Y, to wartość tę wyznaczyć możemy za pomocą następującej zależności:

$$E(Y, R_A) = \int_{-\infty}^{\infty} k_A(y) p(y/\mathbf{x}) dy \qquad (4.2.25)$$

gdzie $k_A(y)$ jest wartością funkcji celu w przypadku alternatywy decyzyjnej A, dla $Y = y$, zaś $p(y/\mathbf{x})$ – funkcją gęstości rozkładu prawdopodobieństwa y, przy danym wzorcu wejściowym $\mathbf{x}$, prognozy zmiennej Y.

Pewnym problemem może być wyznaczenie całki występującej we wzorze (4.2.25). Na szczęście jednak zazwyczaj jest to prosta, jednowymiarowa całka, nie potrzebujemy ponadto jakiejś bardzo precyzyjnej jej wartości, z dokładnością do kilku miejsc po przecinku. W większości przypadków wystarczy nam dosyć przybliżona, żeby nie powiedzieć orientacyjna, wartość naszej funkcji celu, co pozwala na zastosowanie do wyznaczenia wartości (4.2.25) prostych metod całkowania, które bez większego trudu możemy zrealizować choćby w zwykłym arkuszu kalkulacyjnym.

Najprostszy sposób polegać może na wykorzystaniu nieco rozszerzonych metod omawianych w poprzednim punkcie. Dzielimy cały przedział wartości y zmiennej Y, dla których gęstość prawdopodobieństwa $p(y/\mathbf{x})$ jest jeszcze znacząco różna od 0, na szereg podprzedziałów. Na podstawie rozkładu prawdopodobieństwa prognozy zmiennej Y wyznaczamy następnie prawdopodobieństwo każdego podprzedziału jako różnicę wartości dystrybuanty rozkładu na krańcach przedziału (patrz problem 4.2.3). Obliczamy wartość funkcji celu $k_A(y)$ w środku podprzedziału. Oznaczmy ją przez K_A. Zastępujemy wartości funkcji $k_A(y)$ w tym podprzedziale stałą K_A, przybliżając w ten sposób wartości ciągłej zmiennej Y zbiorem wartości dyskretnych. Wyniki alternatywy decyzyjnej opisane są więc obecnie skokową zmienną losową określoną dla każdego podprzedziału. Aby to przybliżenie było dostatecznie dokładne, liczba przedziałów nie może być naturalnie zbyt mała (zależy to od konkretnej sytuacji, ale zazwyczaj wystarczy kilkadziesiąt). By wyznaczyć $E(Y, R_A)$, podobnie jak w problemie 4.2.3, mnożymy prawdopodobieństwa poszczególnych podprzedziałów przez stałe K_A reprezentujące wartości funkcji celu w danym podprzedziale.

Możliwe są oczywiście również inne rozwiązania, w których wykorzystuje się proste kwadratury, takie jak np. złożony wzór trapezów lub kwadraturę Romberga.

Wróćmy wobec tego do sytuacji decyzyjnej z naszego przykładowego problemu 4.2.5 i przyjrzyjmy się wykorzystaniu reguły wartości oczekiwanej wyniku alternatyw decyzyjnych dla rozważanych w nim wariantów zakupu energii elektrycznej. Przypomnijmy, że analizując w rozdziale 3 (a dokładniej mówiąc w podrozdziale 3.4) problem określania rozkładu prawdopodobieństwa krótkoterminowej prognozy zapotrzebowania na energię dla modeli neuronowych i neuronowo-rozmytych, pokazaliśmy, że w przypadkach testowych można było przyjąć założenie o normalnym charakterze tego rozkładu. Obecnie

również więc przyjmiemy takie założenie (podobnie jak w poprzednim punkcie). Pamiętajmy jednak, że kształt rozkładu prawdopodobieństwa powinien zostać sprawdzony dla każdego przypadku poprzez weryfikację rozkładu reszt modelu prognostycznego.

Aby uwzględnić niepewność prognozy, będziemy chcieli wyznaczyć wartości oczekiwane skutków rozważanych alternatyw decyzyjnych, czyli kosztów obu wariantów zakupu energii, przy danym prognozowanym normalnym rozkładzie prawdopodobieństwa jej zapotrzebowania $N(f(\mathbf{x}, \mathbf{w}), \sigma(\mathbf{x}))$. Wartość oczekiwana rozkładu $f(\mathbf{x}, \mathbf{w})$ określona jest przez samą prognozę, czyli przez wyjście sieci neuronowej lub neuronowo-rozmytej, zaś $\sigma(\mathbf{x})$ jest odchyleniem standardowym prognozy. Dla danych z problemu 4.2.5 rozkład prognozy zapotrzebowania na energię będzie więc rozkładem normalnym $N(515, 85)$.

Korzystając z zależności (4.2.25), wyznaczmy zatem wartość oczekiwaną kosztów zakupu energii w wariancie pierwszym $E(ZE, R_1)$:

$$E(ZE, R_1) = \int_{-\infty}^{\infty} k_1(y) p_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(y/\mathbf{x}) dy = \\ = \int_{-\infty}^{\infty} 195 y p_{N(515,85)}(y/\mathbf{x}) dy = 100\,425 \tag{4.2.26}$$

Podobnie wyznaczmy obecnie wartość oczekiwaną kosztów zakupu energii w przypadku wariantu drugiego $E(ZE, R_2)$:

$$E(ZE, R_2) = \int_{-\infty}^{\infty} k_2(y) p_{N(f(\mathbf{x},\mathbf{w}),\sigma(\mathbf{x}))}(y/\mathbf{x}) dy = \\ = \int_{-\infty}^{\infty} \left(185 \min(500, y) + 400 \max(y - 500, 0)\right) p_{N(515,85)}(y/\mathbf{x}) dy = 104\,299 \tag{4.2.27}$$

Spróbujmy przeanalizować otrzymane wyniki. Widzimy, że po uwzględnieniu niepewności prognozy zapotrzebowania spodziewany koszt zakupu energii w wariancie pierwszym wynosi 100 425 złotych i jest niższy niż w przypadku kosztów wariantu drugiego wynoszących 104 299 złotych. W związku z tym, podejmując decyzję, preferowalibyśmy raczej ten wariant zakupu. Zauważmy więc, że jest to zupełnie odwrotna decyzja niż w sytuacji, w której nie uwzględnialiśmy niepewności prognozy, kierując się nią jako faktem pewnym. Porównując wartości otrzymane za pomocą zależności (4.2.23) i (4.2.24), niższe koszty mieliśmy w przypadku wariantu drugiego.

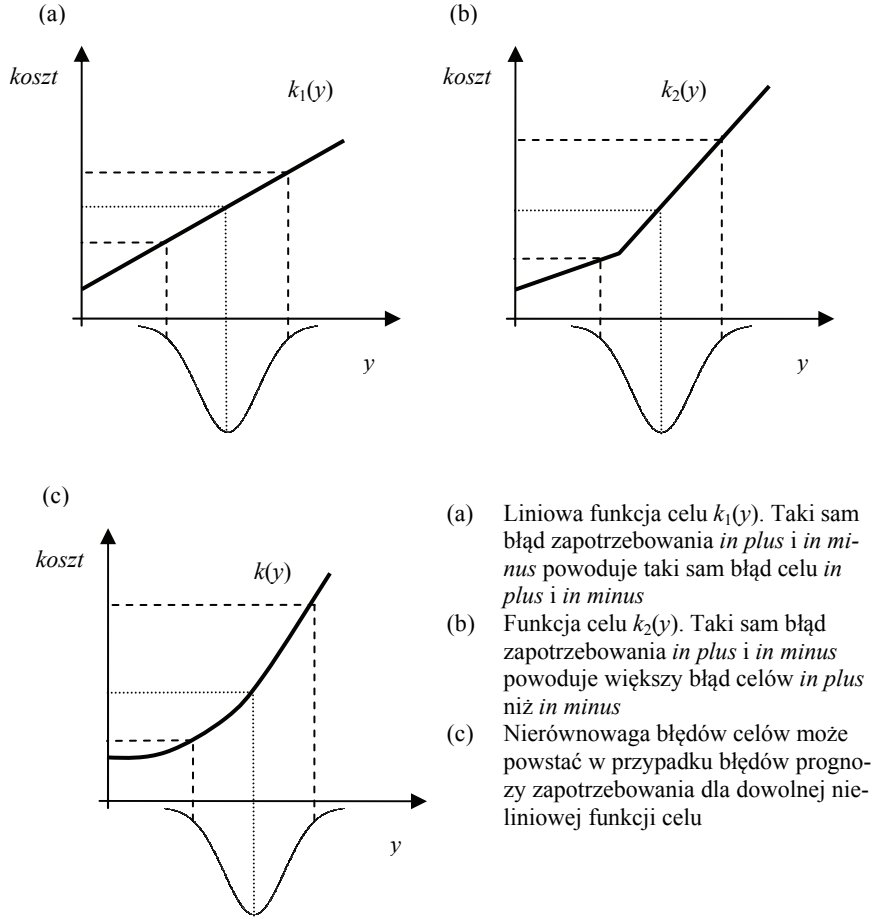
Zauważmy jednak, że wartość oczekiwana kosztów dla wariantu pierwszego, wyznaczona za pomocą zależności (4.2.23) oraz (4.2.26) jest taka sama

i wynosi 100 425 złotych. Natomiast w przypadku wariantu drugiego koszt obliczony przy wykorzystaniu (4.2.27) jest wyraźnie wyższy niż w przypadku (4.2.24). Odpowiednio są to kwoty 104 299 złotych oraz 98 500 złotych. Powstaje więc pytanie, dlaczego uwzględnienie niepewności prognozy zapotrzebowania na energię w jednym przypadku nie powoduje żadnej różnicy w wartości oczekiwanej kosztów, a w drugim przypadku – tak znaczną.

Odpowiedź na to pytanie znaleźć możemy na rysunku 4.2.7. Otóż w przypadku zakupu energii w wariancie pierwszym funkcja kosztów $k_1(y)$ ma charakter liniowy, podobnie jak w punkcie (a) na rysunku 4.2.7. Jednakowa pomyłka w prognozie wartości zapotrzebowania na energię, zarówno o znaku dodatnim jak i ujemnym, dają takie same różnice w kosztach zakupu tej energii. Prognoza określa wartość oczekiwaną zapotrzebowania, nie jest jednak wartością pewną. W praktyce należy się więc spodziewać nieustannych jej odchyleń, które rozkładają się wokół rzeczywistych wartości zapotrzebowania. Ponieważ koszty jednakowych błędów przeszacowania i niedoszacowania prognozy są takie same, w związku z tym koszty zakupu również będą oscylować wokół wartości $k_1(f(\mathbf{x}, \mathbf{w}))$.

Przy zakupie energii w wariancie drugim sytuacja kształtuje się odmiennie. Przypomina ona tę w przypadku (b) na rysunku 4.2.7. Zauważmy, że tutaj pomyłki takiej samej wielkości, ale o przeciwnych znakach, w prognozie zapotrzebowania na energię mogą i często będą dawać zupełnie różne skutki finansowe.

Dokładniej mówiąc, dla przykładowych danych z problemu 4.2.3, jeżeli prognoza jest niedoszacowana i w rzeczywistości, by pokryć zapotrzebowanie, będziemy musieli kupić powiedzmy o 100 MWh więcej niż spodziewane na jej podstawie 515 MWh, to zgodnie z funkcją kosztów $k_2(y)$, określoną zależnością (4.2.21), za tę dodatkową energię zapłacimy 40 000 złotych. Jeżeli natomiast prognoza jest przeszacowana i by pokryć zapotrzebowanie, wystarczy o 100 MWh mniej, to w przykładowych warunkach zapłacimy mniej tylko o 21 725 złotych. Zasadniczo w problemie 4.2.3 koszty pokrycia dodatkowego zapotrzebowania w przypadku niedoszacowania prognozy będą wyższe niż ewentualne oszczędności przy jej przeszacowaniu. Pomimo więc tego, że przeciętny poziom zapotrzebowania na energię jest równy prognozie $f(\mathbf{x}, \mathbf{w})$, przeciętny poziom, wokół którego oscylować będą koszty jej zakupu, okaże się znacznie wyższy niż $k_2(f(\mathbf{x}, \mathbf{w}))$. O ile wyższy? Pozwala to określić wartość oczekiwana kosztów dana przez (4.2.27), która wyważa ponoszone koszty zakupu energii prawdopodobieństwem (gęstością) ich wystąpienia.



(a) Liniowa funkcja celu $k_1(y)$. Taki sam błąd zapotrzebowania *in plus* i *in minus* powoduje taki sam błąd celu *in plus* i *in minus*

(b) Funkcja celu $k_2(y)$. Taki sam błąd zapotrzebowania *in plus* i *in minus* powoduje większy błąd celów *in plus* niż *in minus*

(c) Nierównowaga błędów celów może powstać w przypadku błędów prognozy zapotrzebowania dla dowolnej nieliniowej funkcji celu

Rysunek 4.2.7. Błędy prognozy zapotrzebowania na energię a błędy różnych funkcji celu, które ją wykorzystują
Źródło: opracowanie własne

Podobny problem wystąpić może zawsze w sytuacji, w której funkcja celu oceniająca skutki podejmowanej decyzji ma charakter nieliniowy względem prognozowanej zmiennej w obszarze możliwych jej wartości, to znaczy w obszarze, w którym prawdopodobieństwo tej zmiennej istotnie różni się od zera. Nieliniowość oznacza, że tempo przyrostu lub spadku wartości funkcji celu jest zmienne, w związku z tym skutki błędów *in plus* i *in minus* prognozy, wyrażone w kategoriach naszych celów, mogą być różne. W takim przypadku niezbędne bywa uwzględnienie niepewności prognozy i oszacowanie wartości oczekiwanej funkcji celu w analogiczny sposób do (4.2.27). Pozwala to na ocenę, w jaki sposób niepewność ta wpływa na efekty naszej decyzji.

Gdy funkcje celu mają charakter liniowy, niepewność prognozy dla tego rodzaju zagadnień nie ma znaczenia w tym sensie, że wartość oczekiwaną skutków naszej decyzji możemy wyznaczyć wyłącznie na podstawie prognozy wartości oczekiwanej rozważanej zmiennej.

4.3. Planowanie optymalnej wielkości zakupu w warunkach nierównowagi kosztów nadmiaru i niedoboru energii

W poprzednich dwóch podrozdziałach przeanalizowaliśmy kilka podstawowych zagadnień związanych z podejmowaniem decyzji, których skutki oparte są na zapotrzebowaniu na energię elektryczną w warunkach ryzyka wynikającego z niepewności modelowanego popytu. Rozważania te, pomimo że przydatne w wielu sytuacjach, miały jednak charakter wstępny, a ich celem było raczej przedstawienie podstawowych problemów dotyczących ryzyka decyzji oraz rozmaitych aspektów jego wpływu na wybór alternatywy decyzyjnej. Obecnie przejdziemy do podstawowej części bieżącego rozdziału, mianowicie do zagadnień określania optymalnej wielkości zamówienia zakupu energii elektrycznej w warunkach ryzyka popytowego. Co istotne, **zakładać przy tym będziemy nierównowagę kosztów w przypadku przeszacowania i niedoszacowania wielkości zamówienia w odniesieniu do wielkości rzeczywistego popytu**, to znaczy sytuację, w której koszty zakupu i sprzedaży dodatkowej energii na potrzeby bilansowania są różne.

Zwróćmy uwagę na pewną podstawową różnicę między zagadnieniami, którymi będziemy zajmować się obecnie a tymi rozważanymi poprzednio. W problemach, którymi zajmowaliśmy się wcześniej, dokonywaliśmy wyboru wariantu decyzji spośród dyskretnego i skończonego zbioru rozwiązań. Zadanie określenia optymalnej wielkości zamówienia energii ma jednak charakter ciągły, liczba możliwych decyzji jest nieskończona (a przynajmniej na tyle duża, że musimy ją traktować jako nieskończoną). Jak zobaczymy dalej, zagadnienia tego typu stanowią warianty klasycznego problemu optymalizacji zakupu towaru o ograniczonej trwałości, w warunkach ryzyka popytowego określanego zazwyczaj jako „problem gazeciarza" albo „problem drzewka choinkowego".

4.3.1. Optymalizacja wielkości zakupu przy ograniczonej trwałości towaru w warunkach ryzyka popytowego – klasyczny problem gazeciarza

Specyfika energii elektrycznej jako towaru powoduje konieczność nieustannego równoważenia popytu i podaży. Jak już wspominaliśmy, zdarzenia w systemie elektroenergetycznym rozchodzą się w sposób niemal natychmia-

stowy, z prędkością fali elektromagnetycznej. Reakcja podaży na zmiany popytu musi więc być natychmiastowa i prowadzona na bieżąco. Jeżeli dodamy do tego jeszcze praktyczną niemożność magazynowania energii elektrycznej na skalę przemysłową oraz jej znaczenie jako produktu cywilizacyjnego, niezbędne staje się takie zaprojektowanie mechanizmów handlowych obowiązujących na rynku, aby wymuszały one tego rodzaju równowagę.

Powody te wymuszają organizację konkurencyjnego hurtowego rynku energii elektrycznej w kształcie zbliżonym do prezentowanego w rozdziale 1, z koniecznością grafikowania transakcji przez przedsiębiorstwa handlujące energią oraz z realizacją dostaw pod kontrolą mechanizmów korygujących i rozliczeniowych w postaci rynku bilansującego. Podstawowy cel istnienia tego segmentu rynku polega właśnie na zapewnieniu równowagi istniejącego systemu elektroenergetycznego poprzez umożliwienie w trybie nadzwyczajnym zakupu lub sprzedaży energii w przypadku niezbilansowania popytu i podaży.

Abstrahując nawet od funkcjonowania rynku hurtowego, zasadniczo w kontraktach handlowych na rynku energii elektrycznej trzeba brać pod uwagę ryzyko wynikające z niepewności popytowej. Wspomniany brak możliwości magazynowania energii elektrycznej, a co za tym idzie konieczność stałego bilansowania jej podaży oraz zapotrzebowania, powodują bowiem, że jest ona towarem o ekstremalnie krótkim okresie trwałości. Określanie wielkości zakupu tego typu towarów w warunkach istniejącego ryzyka popytowego stanowi jedno z klasycznych zagadnień zastosowań badań operacyjnych w zarządzaniu, nazywane często problemem gazeciarza lub drzewka choinkowego (Marshall, Oliver 1995). Nietrudno chyba domyślić się pochodzenia tych nazw. Gazety czy choinki świąteczne stanowią typowe przykłady towarów o krótkim okresie trwałości.

Wiele szczegółowych zadań powstających podczas zarządzania wielkością zamówienia na energię elektryczną stanowi określony wariant lub pochodną problemu gazeciarza. W związku z tym przyjrzymy mu się bliżej, rozpoczynając od klasycznego, standardowego sformułowania tego zagadnienia. W naszym przykładzie zmienimy naturalnie rozważany towar z gazet na energię, pozostaniemy jednak w formule płatności za energię zamówioną, bez możliwości rozliczeń niezbilansowania. W specyficznych warunkach obrotu energią elektryczną bardziej przydatny może okazać się wariant problemu omawiany w kolejnym punkcie (problem 4.3.2), oparty na wyborze źródeł i wielkości zakupu z możliwością rozliczeń niezbilansowania przez transakcje z innymi źródłami. Podstawowa wersja problemu gazeciarza również jednak może znaleźć swoje określone zastosowania, zwłaszcza w analizie zagadnień zarządzania stroną popytową, a przede wszystkim aktywną odpowiedzią popytową w sytuacji ryzyka wykorzystywanych prognoz zapotrzebowania.

Zastanówmy się wobec tego nad następującym problemem.

Problem 4.3.1

Przyjmijmy, że przedsiębiorstwo obrotu energią elektryczną planuje jej zakup w postaci krótkoterminowej transakcji z pewnego źródła na rynku, w określonej godzinnej jednostce rozliczeniowej. Przedsiębiorstwo dostarcza energię w cenie r_s, natomiast koszt jednostkowy zakupu z rozważanego źródła wynosi r_z, zakładamy przy tym oczywiście, że $r_s > r_z$. W klasycznym sformułowaniu problemu przyjmujemy następujące założenia:

– przedsiębiorstwo płaci za całość zamówienia, niezależnie od tego, czy sprzeda zamówioną energię czy nie; w związku z tym jeżeli zamówiona zostanie zbyt duża ilość energii, przekraczająca zapotrzebowanie odbiorców, przedsiębiorstwo poniesie straty z powodu niewykorzystania części zamówionej i zapłaconej energii,

– jeżeli zamówienie okaże się zbyt małe w stosunku do zapotrzebowania, zakładamy z kolei, że przedsiębiorstwo nie będzie w stanie dostarczyć odbiorcom całej potrzebnej energii i różnica zostanie pokryta przez innego dostawcę, np. przez operatora systemu dystrybucyjnego; nasze hipotetyczne przedsiębiorstwo poniesie w związku w tym straty z powodu niewykorzystania istniejących szans dodatkowej sprzedaży i uzyskania większego zysku.

Zadanie polega więc na określeniu wielkości zamówienia zakupu z tak, aby zmaksymalizować zysk przedsiębiorstwa ze sprzedaży energii odbiorcom. Możliwe są oczywiście różnego rodzaju warianty problemu gazeciarza, polegające na uwzględnieniu możliwości odzyskania części kwoty za ewentualną zakontraktowaną i zapłaconą, a niesprzedaną energię albo dokupieniu energii w ostatniej chwili po wyższej cenie w przypadku zbyt niskiego zamówienia. Wrócimy jeszcze do tego zagadnienia w przyszłości, w następnym punkcie bieżącego podrozdziału.

Dla danego poziomu zapotrzebowania na energię $ZE = y$ funkcję zysku w naszym zadaniu decyzyjnym możemy zapisać następująco:

$$R(y,z) = r_s \min(y,z) - r_z z \tag{4.3.1}$$

Jej interpretacja jest dosyć oczywista. Jeżeli zamówienie będzie mniejsze od zapotrzebowania, przedsiębiorstwo sprzeda odbiorcom jedynie tyle energii, ile zamówiło. Gdy zamówienie będzie zbyt duże, sprzedane zostanie i tak jedynie tyle energii, ile wynosi zapotrzebowanie. W każdym zaś przypadku przedsiębiorstwo płaci za całą zakontraktowaną energię.

Zamówienia energii dokonujemy naturalnie, nie znając jeszcze rzeczywistej wielkości zapotrzebowania. Zakładamy natomiast, że znamy jego prognozowany rozkład prawdopodobieństwa, określony za pomocą funkcji gęstości rozkładu $p(y)$ lub dystrybuanty $P(y)$, który może zostać wykorzystany do oceny ryzyka

skutków decyzji odnośnie do wielkości zamówienia. Dla różnych wartości zapotrzebowania $ZE = y$, jakie zrealizują się faktycznie z rozkładu $p(y)$, skutki decyzji mierzone zyskiem przedsiębiorstwa $R(y, z)$, określonym przez (4.3.1), będą odmienne.

Gdy skutki decyzji obarczone są niepewnością opisywaną przez pewien rozkład prawdopodobieństwa, to zgodnie z rozważaniami w punkcie 4.2, obiektywnym, wolnym od określonego nastawienia do ryzyka kryterium wyboru alternatywy decyzyjnej jest reguła optymalizacji wartości oczekiwanej tych skutków. Zauważmy jednak, że w obecnym przypadku możliwych alternatyw decyzyjnych, czyli wielkości zamówienia z, jest nieskończenie wiele (a przynajmniej na tyle dużo, że musimy w ten sposób potraktować całą tę sytuację). Nie możemy więc, tak jak to robiliśmy wcześniej w punkcie 4.2, wyznaczyć wartości oczekiwanej zysku ze sprzedaży energii enumeratywnie, dla każdego z, a następnie wybrać najlepszego wariantu.

Zamiast tego musimy zastosować podejście optymalizacyjne, które polega na znalezieniu wielkości zamówienia z maksymalizującego wartość oczekiwaną zysku ze sprzedaży energii $E(R(ZE, z))$ dla różnych możliwych wartości zapotrzebowania odbiorców y, czyli względem rozkładu prawdopodobieństwa tego zapotrzebowania ZE. Nasze kryterium oceny skutków decyzji określimy więc przez:

$$E(R(ZE,z)) = \int_0^{\infty} \left(r_s \min(y,z) - r_z z\right) p(y)\,dy \tag{4.3.2}$$

Całka w (4.3.2) obliczana jest od 0 do ∞, ponieważ wielkość zamówienia z musi oczywiście być liczbą nieujemną.

Należy więc wybrać taką wielkość zamówienia zakupu energii z^*, $z^* \geq 0$, dla której wartość oczekiwana zysku (4.3.2) będzie jak największa. Zauważmy, że bezpośrednia maksymalizacja funkcji celu (4.3.2) może być niewygodna z powodu występującej w niej operacji minimum. Dlatego więc często korzysta się z równoważnej postaci wartości oczekiwanej zysku $E(R(ZE, z))$, w której zamiast gęstości rozkładu zapotrzebowania na energię $p(y)$, stosuje się jego dystrybuantę $P(y)$.

Skorzystajmy tutaj z pewnej właściwości wartości oczekiwanej. Otóż dla dowolnej nieujemnej zmiennej losowej T, o gęstości rozkładu prawdopodobieństwa $p_T(t)$ oraz dystrybuancie tego rozkładu $P_T(t)$, wartość oczekiwaną zmiennej T możemy zapisać w następującej postaci:

$$E(T) = \int_0^{\infty} (1 - P_T(t))\,dt \tag{4.3.3}$$

pod warunkiem, że odpowiednie całki istnieją.

Zależność (4.3.3) wynika niemal natychmiast z faktu, że, zgodnie z definicją, dystrybuanta jest funkcją pierwotną funkcji gęstości prawdopodobieństwa. Jeżeli więc teraz do wyznaczenia wartości oczekiwanej $E(T)$ zastosujemy wzór na całkowanie przez części oraz weźmiemy następnie pod uwagę fakt, że dla $t \to \infty$ dystrybuanta $P_T(t) \to 1$, zaś dla $t \to 0$ dystrybuanta $P_T(t) \to 0$, to rozumiejąc niektóre z poniższych przekształceń w sensie granicznym, otrzymamy następującą zależność:

$$\begin{aligned} &\int_0^\infty (1 - P_T(t))\,dt = \left[t(1 - P_T(t))\right]_0^\infty - \int_0^\infty -tp_T(t)\,dt = \\ &= \int_0^\infty tp_T(t)\,dt = E(T) \end{aligned} \tag{4.3.4}$$

Jeżeli teraz zmienna losowa T zdefiniowana jest jako $T = \min(Y, z)$, gdzie Y jest również nieujemną zmienną losową, taką jak interesujące nas zapotrzebowanie na energię, zaś z dodatnią stałą, np. wielkością zamówienia, to wówczas mamy

$$E(T) = \int_0^\infty (1 - P_{\min(Y,z)}(t))\,dt = \int_0^z (1 - P_Y(t))\,dt + \int_z^\infty (1 - P_z(t))\,dt \tag{4.3.5}$$

Zauważmy jednak, że dla $t \geq z$, $P_z(t) = \Pr(z \leq t) = 1$, więc druga całka w (4.3.5) jest równa 0. Ostatecznie otrzymujemy więc:

$$E(T) = \int_0^z (1 - P_Y(t))\,dt \tag{4.3.6}$$

Wróćmy wobec tego do rozważanego przez nas problemu 4.3.1 oraz wartości oczekiwanej zysku $E(R(ZE, z))$, przy danej wielkości zamówienia z i zapotrzebowaniu na energię elektryczną ZE. Korzystając z addytywności całki w zależności (4.3.2) oraz biorąc pod uwagę fakt, że całka z funkcji gęstości prawdopodobieństwa w całej przestrzeni jest równa 1, możemy napisać:

$$\begin{aligned} E(R(ZE,z)) &= r_s \int_0^\infty \min(y,z)\,p(y)\,dy - r_z z \int_0^\infty p(y)\,dy = \\ &= r_s \int_0^\infty \min(y,z)\,p(y)\,dy - r_z z \end{aligned} \tag{4.3.7}$$

Korzystając teraz z zależności (4.3.6), otrzymujemy:

$$E(R(ZE,z)) = r_s \int_0^z (1-P(y))dy - r_z z = \\ = r_s \int_0^z dy - r_s \int_0^z P(y)dy - r_z z = r_s z - r_s \int_0^z P(y)dy - r_z z \tag{4.3.8}$$

Ostatecznie więc, porządkując nieco wyznaczoną zależność (4.3.8), otrzymujemy finalną postać funkcji wartości oczekiwanej zysku $E(R(ZE, z))$, w której wykorzystuje się dystrybuantę prognozowanego rozkładu prawdopodobieństwa zapotrzebowania na energię elektryczną $P(y)$. Możemy ją zapisać w następujący sposób:

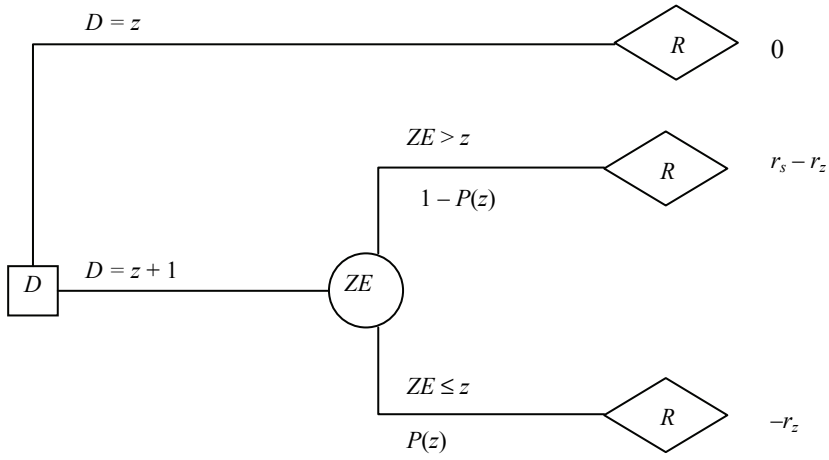
$$E(R(ZE,z)) = (r_s - r_z)z - r_s \int_0^z P(y)dy \tag{4.3.9}$$

Podsumowując więc naszą dotychczasową dyskusję w bieżącym punkcie, należy stwierdzić, że optymalna wielkość zamówienia z energii elektrycznej w problemie 4.3.1 ma minimalizować ryzyko popytowe wynikające z niepewności prognozowanego zapotrzebowania. Z jednej strony, szacując tę wielkość, trzeba brać pod uwagę ryzyko przeszacowania popytu i zakontraktowania zbyt dużej ilości energii, która nie zostanie sprzedana, z drugiej – ryzyko jego niedoszacowania i niemożności pełnego zaspokojenia potrzeb odbiorców ze źródeł przedsiębiorstwa. Elementy te bierzemy pod uwagę w funkcji zysku (4.3.1). Należy więc wybrać taką wielkość zamówienia z, która maksymalizuje wartość oczekiwaną funkcji zysku dla prognozowanego rozkładu prawdopodobieństwa zapotrzebowania na energię ZE określoną jedną z równoważnych zależności (4.3.2) lub (4.3.9).

Określenie optymalnej wielkości zamówienia z^* polega więc na rozwiązaniu odpowiedniego zagadnienia optymalizacji stochastycznej w warunkach ryzyka popytowego. Maksymalizacja (4.3.2) lub (4.3.9) względem z w takim przypadku nie jest zadaniem trudnym i można ją przeprowadzić na wiele sposobów. Najprostszy z nich stanowi chyba zastosowanie metody analizy krańcowej albo, ściślej mówiąc, analizy przyrostów krańcowych, która polega na zastąpieniu ciągłego problemu optymalizacyjnego równaniem różnicowym, wynikającym z badania zmian wartości oczekiwanych zysków dla niewielkich (jednostkowych) zmian wielkości zamówienia. Metoda ta jest w naszym przypadku uprawniona, ponieważ pamiętać należy, że wielkość zamówienia stanowi zmienną *quasi*-ciągłą, jako że handel energią odbywa się w określonych jednostkach

handlowych. Jest ona przy tym dosyć ilustracyjna, więc będziemy ją szeroko stosować w kilku kolejnych zagadnieniach prezentowanych w dalszej części bieżącego rozdziału.

Zastosowanie analizy krańcowej do problemu 4.3.1 polega na porównaniu, jak zmienia się nasza funkcja zysku, jeżeli zwiększamy wielkość zamówienia zakupu energii elektrycznej o jeden w stosunku do pewnego poziomu podstawowego z. Sprowadzamy więc zagadnienie do znanego nam już z punku 4.2 schematu dwóch alternatyw decyzyjnych – bezpiecznej (wolnej od ryzyka) oraz ryzykownej. Alternatywa bezpieczna polega, rzecz jasna, na pozostawieniu wielkości zamówienia na poziomie z. Rozwiązanie drugie, w którym badamy konsekwencje zakupu dodatkowej jednostki energii, ma charakter ryzykowny, ponieważ zmiana zysku związanego z tym zakupem w stosunku do poprzedniego poziomu jest obarczona niepewnością i może dać różne skutki w zależności od wartości zapotrzebowania.



Rysunek 4.3.1. Drzewo decyzyjne analizy krańcowej dla zadania określania optymalnej wielkości zamówienia w problemie 4.3.1
Źródło: opracowanie własne

Przedstawiana sytuacja dla rozważanego przez nas problemu 4.3.1 została zobrazowana w postaci drzewa decyzyjnego na rysunku 4.3.1. Jak widzimy, biorąc jako punkt odniesienia zysk ze sprzedaży otrzymywany dla wielkości zamówienia zakupu energii z, zwiększenie zamówienia o 1 daje dwie różne możliwości, jeśli chodzi o zmianę zysku, w zależności od faktycznego poziomu zapotrzebowania na energię ZE:

– jeżeli zapotrzebowanie na energię okaże się mniejsze od wielkości zamówienia z bądź mu równe (dolna gałąź w węźle ZE), to dodatkowa zakupiona

jednostka energii nie będzie mogła być sprzedana; możemy mówić więc o porażce, ponieważ zwiększenie wielkości zamówienia o 1 spowoduje obniżenie wartości zysku o koszt jej zakupu $-r_z$; prawdopodobieństwo takiego skutku wyboru ryzykownej alternatywy decyzyjnej równe jest $\Pr(ZE \le z)$, czyli dystrybuancie spodziewanego rozkładu prawdopodobieństwa zapotrzebowania na energię $P(z)$,

– jeżeli zapotrzebowanie na energię będzie większe od wielkości zamówienia z (górna gałąź w węźle ZE), to znaczy, że zapotrzebowanie to musi wynosić co najmniej $z + 1$, czyli dodatkowa zamówiona jednostka energii może jeszcze zostać sprzedana; zwiększenie wielkości zamówienia o 1 kończy się więc w tym przypadku sukcesem, powodując powiększenie wartości zysku o różnicę między ceną sprzedaży i zakupu tej jednostki energii $r_s - r_z$; prawdopodobieństwo takiego stanu rzeczy wynosi z kolei $1 - \Pr(ZE \le z)$, czyli $1 - P(z)$.

Jak już mówiliśmy w punkcie 4.2.1, ponieważ decyzja o zwiększeniu zamówienia z o 1 może zaowocować różnymi skutkami, to jako obiektywną miarę oceny alternatyw decyzyjnych przyjmiemy wartości oczekiwane tychże skutków, czyli w naszym wypadku $E(R(ZE, z))$ oraz $E(R(ZE, z+1))$. Oczywiście decyzja bezpieczna $D = z$ daje stały zysk równy 0, więc wartość oczekiwana zysku $E(R(ZE, z))$ jest również równa 0. Natomiast w przypadku opcji ryzykownej wartość oczekiwaną zysku dla zwiększonego zamówienia $z + 1$ wyznaczymy, ważąc poszczególne skutki tego wariantu decyzji prawdopodobieństwami ich realizacji. W naszym przypadku, analizując sytuację przedstawioną na drzewie decyzyjnym na rysunku 4.3.1, możemy więc zapisać:

$$E(R(ZE, z+1)) = (r_s - r_z)(1 - P(z)) - r_z P(z) \tag{4.3.10}$$

Nietrudno dalej zauważyć, że optymalna wielkość zamówionej energii z^*, maksymalizująca wartość oczekiwaną zysku z tego zamówienia $E(R(ZE, z))$ (określoną przez zależność (4.3.2) lub (4.3.9)), przy niepewnym popycie będzie równa wartości z^* stanowiącej rozwiązanie równania różnicowego:

$$E(R(ZE, z+1)) - E(R(ZE, z)) = 0 \tag{4.3.11}$$

Różnicę po lewej stronie równania (4.3.11) określamy jako margines wartości oczekiwanej zysku. Dla małych wielkości zamówienia z, dużo niższych od wartości prognozy popytu zapotrzebowania na energię ZE, przyrost ten będzie dodatni. Wynika to z zależności (4.3.10) i analizy drzewa decyzyjnego naszego problemu na rysunku 4.3.1. Zauważmy bowiem, że zwiększenie wielkości zamówienia energii z o jedną jednostkę w przypadku sukcesu (sprzedaży dodatkowej jednostki energii) daje pewien dodatni zysk $r_s - r_z$, natomiast

w przypadku porażki skutkuje wartością ujemną, czyli stratą $-r_z$. W sumie ważonej (4.3.10) mnożymy więc pewną wartość dodatnią przez prawdopodobieństwo sukcesu, zaś pewną wartość ujemną przez prawdopodobieństwo porażki.

Jeżeli z jest małe, to zwiększenie zamówienia o 1 z dużym prawdopodobieństwem będzie skutkować sprzedażą dodatkowej jednostki i zyskiem $r_s - r_z$. Prawdopodobieństwo, że dodatkowa jednostka energii nie zostanie sprzedana i poniesiemy stratę $-r_z$ jest niewielkie. W związku z tym w (4.3.10) mnożymy ujemny element przez bardzo małe prawdopodobieństwo, bliskie wartości 0, zaś dodatni – przez duże, bliskie 1. Wartość oczekiwana zysku dla zwiększonej wielkości zamówienia $E(R(ZE, z+1))$, a co za tym idzie, margines wartości oczekiwanej zysku będą dodatnie.

Ponieważ przyrost wartości oczekiwanej zysku przy wzroście wielkości zamówienia jest dodatni, opłaca się zamówienie zwiększyć. Musimy jednak zadać sobie pytanie, czy tak będzie zawsze. Z pewnością nie. Wraz ze wzrostem wielkości zamówienia z maleje bowiem prawdopodobieństwo sukcesu, czyli sprzedaży jeszcze jednej, dodatkowej jednostki energii, a rośnie prawdopodobieństwo, że nie zostanie ona sprzedana. Przyrost wartości oczekiwanej $E(R(ZE, z+1)) - E(R(ZE, z))$ będzie więc coraz mniejszy, ponieważ w zależności (4.3.10) maleje prawdopodobieństwo członu dodatniego związanego z sukcesem, natomiast rośnie prawdopodobieństwo członu ujemnego. Dla pewnej wartości z zmiana ta stanie się w końcu mniejsza od zera. W tej sytuacji dalsze zwiększanie wielkości zamówienia energii powodowałoby spadek spodziewanej wartości zysku. W związku z tym nie powinniśmy tego robić.

Widzimy więc, że zwiększanie wielkości zamówienia opłaca się, kiedy przyrost wartości oczekiwanej zysku jest dodatni, zaś przestaje się opłacać z chwilą, gdy stanie się on ujemny. Optymalna wielkość zamówienia z^* wyznaczana jest wobec tego przez punkt równowagi, to jest taką wartość z, dla której przyrost wartości oczekiwanej jest równy 0, czyli dla rozwiązania równania (4.3.11).

Rozwiązanie równania (4.3.11) stanowi dosyć proste zadanie. Pamiętając, że wartość oczekiwana zysku dla wielkości zamówienia energii równej z jest w naszym przypadku poziomem odniesienia zysku równym 0, tzn. $E(R(ZE, z)) = 0$, oraz podstawiając $E(R(ZE, z+1))$ z zależności (4.3.10), otrzymujemy:

$$(r_s - r_z)(1 - P(z)) - r_z P(z) = 0 \tag{4.3.12}$$

Wykonując proste przekształcenia algebraiczne, wyznaczamy rozwiązanie (4.3.12), względem $P(z)$:

$$
\begin{aligned}
&(r_s - r_z) - P(z)(r_s - r_z) - r_z P(z) = 0 \\
&r_s - r_z - P(z)r_s + P(z)r_z - r_z P(z) = 0 \\
&r_s - r_z - P(z)r_s = 0 \\
&P(z)r_s = r_s - r_z
\end{aligned}
\tag{4.3.13}
$$

skąd, ostatecznie, otrzymujemy dobrze znany wynik wspomnianego wcześniej, standardowego problemu gazeciarza:

$$P(z^*) = \frac{r_s - r_z}{r_s} \tag{4.3.14}$$

Czasami jest on przedstawiany w alternatywnej postaci:

$$1 - P(z^*) = \frac{r_z}{r_s} \tag{4.3.14a}$$

W naszym stosunkowo prostym problemie zależność (4.3.14) możemy naturalnie otrzymać w inny, bardziej klasyczny sposób. Jeżeli weźmiemy naszą funkcję celu, czyli wartość oczekiwaną funkcji zysku ze sprzedaży energii (4.3.9) i wyznaczymy jej pochodną względem wielkości zamówienia z, otrzymamy:

$$
\begin{aligned}
&\frac{\partial}{\partial z} E(R(ZE, z)) = \frac{\partial}{\partial z}\left((r_s - r_z)z - r_s \int_0^z P(y)dy \right) = \\
&= (r_s - r_z) - r_s P(z)
\end{aligned}
\tag{4.3.15}
$$

Aby określić optymalną wartość z^*, $z^* \geq 0$, maksymalizującą spodziewane zyski, przyrównujemy pochodną funkcji celu (4.3.15) do zera:

$$\frac{\partial}{\partial z} E(R(ZE, z)) = (r_s - r_z) - r_s P(z) = 0 \tag{4.3.16}$$

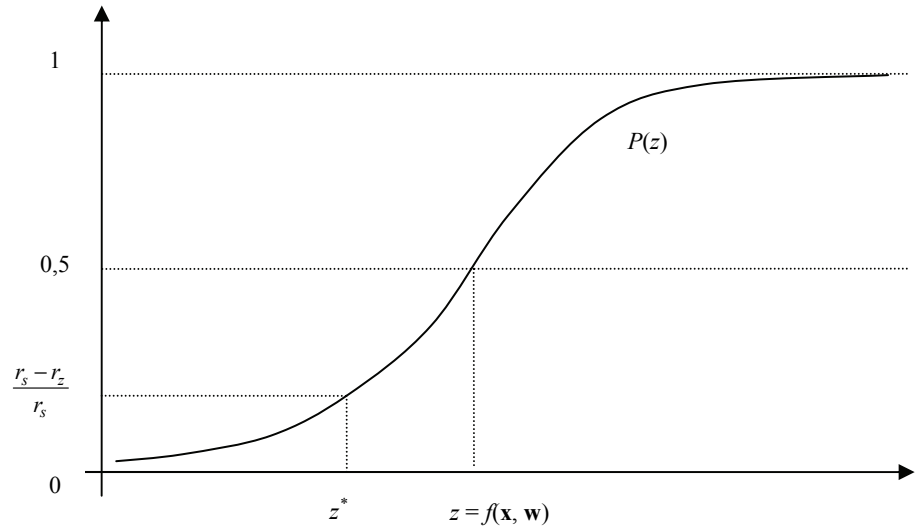
Rozwiązanie równania (4.3.16) jest w zasadzie trywialne. Po kilku prostych przekształceniach algebraicznych otrzymamy zależność określającą optymalną wartość z^* w postaci formuły (4.3.14) lub alternatywnie (4.3.14a).

Jak więc widzimy, w dosyć powszechnej sytuacji zdefiniowanej w problemie 4.3.1 optymalna wielkość zamówienia energii elektrycznej z^* określana jest przez (4.3.14), to znaczy wyznaczona zostaje przez wartość dystrybuanty $P(z)$ przewidywanego rozkładu prawdopodobieństwa zapotrzebowania na energię, odpowiadającą stosunkowi marży ze sprzedaży energii do ceny jej sprzedaży (lub alternatywnie z^* określane jest przez (4.3.14a), czyli przez taką wartość dystrybuanty zapotrzebowania na energię, że $1 - P(z)$ jest równe stosunkowi kosztu jednostkowego zakupu do ceny sprzedaży).

Zwróćmy uwagę na interesujące konsekwencje dla wykorzystania prognoz krótkoterminowego zapotrzebowania na energię w procesie planowania wielkości zamówienia na rynku energii. Zazwyczaj informacją, na podstawie której podejmuje się decyzję o wolumenie zamówienia, jest sama wartość prognozy zapotrzebowania. Tymczasem, jak wynika z (4.3.14), taki sposób postępowania na ogół ma charakter nieoptymalny. Modele prognostyczne krótkoterminowego zapotrzebowania na energię budowane są na podstawie kryteriów minimalizacji błędu statystycznego (w większości przypadków kwadratowego) prognozy. Jak dyskutowaliśmy w rozdziale 3.1.1, samą prognozę, czyli wyjście modelu prognostycznego $f(\mathbf{x}, \mathbf{w})$, możemy wówczas interpretować jako przybliżenie wartości oczekiwanej $E(ZE / \mathbf{x})$ przewidywanego rozkładu prawdopodobieństwa zapotrzebowania na energię ZE, przy określonej wartości zmiennych wejściowych modelu $\mathbf{x}$.

Tymczasem, jak wynika z (4.3.14), wartość oczekiwana rozkładu prawdopodobieństwa zapotrzebowania, $f(\mathbf{x}, \mathbf{w})$, bardzo rzadko będzie odpowiadać optymalnej wielkości zamówienia energii elektrycznej z^*. Może stać się tak wyłącznie w przypadku, gdy różnica między ceną sprzedaży a kosztem zakupu $r_s - r_z$ stanowić będzie połowę ceny sprzedaży r_s. A ponadto rozkład prawdopodobieństwa zapotrzebowania musi mieć charakter symetryczny, w przeciwnym razie optymalną wielkość zamówienia energii wyznaczać będzie mediana, a nie wartość oczekiwana rozkładu.

Dobrą ilustrację tego problemu stanowi rysunek 4.3.2. Widzimy na nim od razu, dlaczego optymalna wielkość zamówienia energii dla problemu 4.3.1 różni się zazwyczaj od wartości prognozy otrzymanej przy użyciu modelu statystycznego określającego wartość oczekiwaną procesu zapotrzebowania na energię (zakładamy przy tym, że rozkład ten jest symetryczny i mediana odpowiada wartości oczekiwanej). Otóż jeżeli różnica ceny sprzedaży i kosztów zakupu $r_s - r_z$ stanowi jedynie niewielką część ceny sprzedaży r_s (tak jak to widzimy na rysunku 4.3.2), to optymalna wielkość zamówienia powinna być niższa od wartości prognozy, ponieważ w tej sytuacji mamy $r_s - r_z < r_z$.



Rysunek 4.3.2. Ilustracja graficzna rozwiązania zagadnienia optymalnej wielkości zamówienia zakupu energii w warunkach ryzyka prognozy w problemie 4.3.1
Źródło: opracowanie własne

Jeżeli głębiej zastanowimy się nad tym zagadnieniem, to wniosek ten powinien być zresztą dosyć oczywisty. Prognozy wartości oczekiwanej oparte są na minimalizacji błędu kwadratowego modelu i jednakowo traktują odchylenia dodatnie i ujemne. Wartość oczekiwana zapotrzebowania na energię jest jedynie wartością najbardziej prawdopodobną, przeciętną. Faktyczne zapotrzebowanie niemal nigdy nie zrealizuje się dokładnie na poziomie prognozy. W sposób naturalny będzie ono wokół niej oscylować, skutkując dodatnimi i ujemnymi odchyleniami.

Przyjmując więc wielkość zamówienia na poziomie wartości oczekiwanej zapotrzebowania na energię (prognozy) z jednakowym prawdopodobieństwem narażamy się na ryzyko przeszacowania bądź niedoszacowania zamówienia. Zwróćmy jednak uwagę, że w przypadku zamówienia zbyt małej ilości energii spodziewane zyski zmniejszają się tylko o różnicę między jej ceną sprzedaży a kosztem jednostkowym zakupu $r_s - r_z$ na każdej jednostce. Natomiast przy przeszacowaniu wielkości zamówienia o takim samym rozmiarze w kategoriach bezwzględnych ryzykujemy obniżeniem zysku o r_z na każdej jednostce energii, która pozostanie niesprzedana, a to w sytuacji przedstawionej na rysunku 4.3.2 stanowi kwotę znacznie wyższą.

Mamy więc do czynienia z nierównomiernymi kosztami finansowymi takiego samego odchylenia prognozy na plus i na minus, czego w modelach statystycznych prognozujących wartość oczekiwaną zapotrzebowania na energię

elektryczną w ogóle nie bierze się pod uwagę. W sytuacji opisanej na rysunku 4.3.2 maksymalizująca oczekiwane zyski optymalna wielkość zamówienia z^*, aby skompensować większe ryzyko przeszacowania prognozy, musi być niższa od przewidywanego przez model prognostyczny przeciętnego poziomu zapotrzebowania na energię. Może częściej zamówienie okaże się za małe i moglibyśmy sprzedać więcej, ale to i tak będzie nam się bardziej opłacać, niż gdybyśmy mieli towaru za dużo. Nie należy oczywiście przesadzać. Taka sytuacja nie może być zbyt częsta. Wielkość zamówienia powinna być niższa od prognozy – o tym mówi nam właśnie zależność (4.3.14) (lub (4.3.14a)).

Z drugiej strony, gdy różnica ceny sprzedaży i kosztów zakupu jednostki energii $r_s - r_z$ stanowi większą część ceny sprzedaży r_s (przekraczającą jej połowę), to wynikająca z formuły (4.3.14) optymalna wielkość zamówienia z^* będzie naturalnie wyższa od prognozowanej przez model wartości oczekiwanej rozkładu prawdopodobieństwa zapotrzebowania $f(\mathbf{x}, \mathbf{w})$. Widać to wyraźnie na rysunku 4.3.2. I znów jest to zgodne z logiką. W takim bowiem przypadku różnica ceny sprzedaży i kosztów zakupu $r_s - r_z$ jest większa od r_z, a więc niedoszacowane zamówienie przynosi większą redukcję spodziewanego zysku niż to przeszacowane. W tej sytuacji opłaca się więc zamówić więcej niż wynosi prognoza.

Analizując w rozdziale 3 problem modelowania niepewności prognozy krótkoterminowego zapotrzebowania na energię elektryczną przy użyciu badanych modeli neuronowych i neuronowo-rozmytych, wskazywaliśmy, że rozkład prawdopodobieństwa prognozy popytu powinien mieć charakter rozkładu normalnego Gaussa. Oczywiście w każdym konkretnym przypadku, jak wskazywaliśmy wcześniej, założenie to wymaga weryfikacji empirycznej, ale w badanych modelach wyniki dosyć jednoznacznie pozwalały potwierdzić hipotezę gaussowską.

W przypadku rozkładu normalnego prognozy zapotrzebowania na energię elektryczną problem określenia optymalnej wielkości zamówienia sprowadza się do wyznaczenia prostej poprawki dla prognozy otrzymywanej na wyjściu sieci neuronowej lub neuronowo-rozmytej. Dystrybuanta rozkładu normalnego jest funkcją różnowartościową. Przekształcając zatem (4.3.14), optymalną wartość z^* możemy wyznaczyć jako:

$$z^* = P^{-1}\left(\frac{r_s - r_z}{r_s}\right) \tag{4.3.17}$$

gdzie P^{-1} stanowi funkcję odwrotną do dystrybuanty (czyli jest po prostu kwantylem) prognozowanego rozkładu prawdopodobieństwa zapotrzebowania na energię elektryczną $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$, gdzie $f(\mathbf{x}, \mathbf{w})$ jest wyjściem modelu prognostycznego (prognozą), zaś $\sigma_y(\mathbf{x})$ odchyleniem standardowym prognozy.

W rozdziale 3 szeroko omawialiśmy metody szacowania odchylenia standardowego wyjścia modelu dla badanych rodzajów sieci neuronowych i neuronowo-rozmytych.

Zależność (4.3.17) możemy zapisać w równoważnej postaci, normalizując rozkład prawdopodobieństwa zapotrzebowania na energię $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$ do standardowego rozkładu normalnego $N(0, 1)$. Optymalną wielkość zamówienia energii elektrycznej z^* wyznaczamy wówczas jako prostą korektę do prognozy otrzymanej z modelu neuronowego lub neuronowo-rozmytego:

$$z^* = f(\mathbf{x}, \mathbf{w}) + Q_{N(0,1)}\left(\frac{r_s - r_z}{r_s}\right) \cdot \sigma_y(\mathbf{x}) \tag{4.3.18}$$

gdzie $Q_{N(0,1)}(\cdot)$ jest kwantylem rozkładu normalnego standardowego $N(0, 1)$, natomiast $f(\mathbf{x}, \mathbf{w})$ i $\sigma_y(\mathbf{x})$, tak samo jak w przypadku (4.3.17), są wyjściem sieci neuronowej (neuronowo-rozmytej) i jego odchyleniem standardowym wyznaczonym za pomocą metod opisanych w rozdziale 3.

4.3.2. Optymalna wielkość zakupu energii elektrycznej na rynku w warunkach ryzyka popytowego

Sytuacja analizowana w poprzednim punkcie, w problemie 4.3.1, w pewnych warunkach, może oczywiście występować w obrocie energią elektryczną, zwłaszcza w powiązaniu z modelowaniem aktywnej odpowiedzi popytowej. Tym niemniej należy zwrócić uwagę, że niektóre warunki określone w tym problemie występują dosyć rzadko. Na rynkach energii elektrycznej bardziej powszechna jest sytuacja, w której dostawca energii zarządza ewentualnym niezbilansowaniem swoich odbiorców, korzystając z dostępnych mechanizmów, takich jak transakcje na giełdach energii o krótkim horyzoncie czasowym czy też na rynku bilansującym.

Zastanówmy się w związku z tym nad następującym problemem.

Problem 4.3.2

Przyjmijmy, podobnie jak w problemie 4.3.1, że przedsiębiorstwo obrotu energią elektryczną planuje jej zakup w postaci krótkoterminowej transakcji z pewnego źródła na rynku, w określonej godzinnej jednostce rozliczeniowej. Przedsiębiorstwo dostarcza energię w cenie r_s, natomiast koszt jednostkowy zakupu z rozważanego źródła wynosi r_z; zakładamy przy tym oczywiście, że $r_s > r_z$. W odróżnieniu od problemu sformułowanego w poprzednim punkcie obecnie przyjmujemy następujące założenia.

Przedsiębiorstwo płaci za całość zamówienia, niezależnie od tego, czy sprzeda zamówioną energię czy nie. Jeżeli jednak zamówiona zostanie zbyt duża

ilość energii, która przekracza zapotrzebowanie odbiorców, przedsiębiorstwo dla zbilansowania zamówienia i faktycznego popytu może sprzedać posiadaną nadwyżkę, uzyskując przychód z tej sprzedaży wynoszący r_r złotych na każdej jednostce energii. Zakładamy przy tym, że „ratunkowy" jednostkowy przychód ze sprzedaży nadwyżki jest niższy od normalnego kosztu jednostkowego zakupu, tj. $r_r < r_z$, więc energia ta zbywana jest z pewną stratą i w ten sposób odzyskujemy jedynie część kosztów z nią związanych.

W sytuacji, w której zamówienie będzie zbyt małe w stosunku do zapotrzebowania, przedsiębiorstwo, by pokryć zapotrzebowanie swoich odbiorców, musi kupić energię z innego źródła, na przykład od operatora systemu dystrybucyjnego albo na rynku bilansującym. Koszty zakupu tego rodzaju energii wynoszą r_i złotych dla każdej jednostki niedoboru. W tym przypadku jednak koszt jednostkowy zakupu energii dodatkowej jest wyższy od jednostkowego kosztu zakupu w normalnych warunkach, a więc $r_z < r_i$.

Zwróćmy uwagę, że sytuacja przedstawiona w problemie 4.3.2 nie istnieje w tej chwili na rynku polskim, a przynajmniej nie w formie bezpośredniej. Zgodnie z obecnie obowiązującymi zasadami ceny zakupu (CROz) i sprzedaży (CROs) energii przez rynek bilansujący w Polsce są takie same. Warunki zbliżone do opisanych obowiązywały jednak na wcześniejszych etapach działania rynku energii. Rozchylenie cen CROz i CROs rozważane jest również jako jedno z rozwiązań na przyszłość. Podobne problemy pojawią się także przy planowanym przejściu rozliczeń na system tzw. cen węzłowych. Ponadto do zbliżonych sytuacji decyzyjnych może dochodzić również w warunkach obecnych w wyniku optymalizacji źródeł zakupowych na rynkach o różnych wyprzedzeniach czasowych.

Zadanie polega więc, tak samo jak w problemie 4.3.1, na określeniu wielkości zamówienia zakupu $z = z^*$, $z \geq 0$, tak aby zmaksymalizować zysk przedsiębiorstwa ze sprzedaży energii. Dokładniej rzecz biorąc, jak pokażemy niżej, w sytuacji opisanej w bieżącym przykładzie, zamiast szukać wielkości zamówienia maksymalizującej spodziewane zyski, wygodniej będzie nam przyjąć wariant optymalizacji kosztowej.

Funkcja zysku dla naszego zadania decyzyjnego przy danym poziomie zapotrzebowania na energię $ZE = y$ przyjmuje postać:

$$R(y,z) = r_s y - \left(r_z z - \max(z-y,0)r_r + \max(y-z,0)r_i\right) \tag{4.3.19}$$

Również i w tym przypadku interpretacja zależności (4.3.19) jest dosyć oczywista. Pierwszy człon $r_s y$ określa przychód ze sprzedaży energii. Przedsiębiorstwo sprzedaje odbiorcom zawsze tyle energii, ile wynosi popyt. Dalej – przedsiębiorstwo płaci za całą zamówioną energię $r_z z$. Jeżeli zamówienie jest przeszacowane, czyli występuje nadmiar zamówionej energii ponad popyt,

można go odsprzedać, odzyskując część kosztów wydanych na niepotrzebnie zamówioną energię, w kwocie $\max(z - y,0)\, r_r$. Jeżeli natomiast zamówienie jest niedoszacowane, czyli zapotrzebowanie na energię okazuje się wyższe od zakontraktowanej wielkości, brakującą energię przedsiębiorstwo musi dokupić, ponosząc dodatkowy koszt w wysokości $\max(y - z,0)\, r_i$.

Zauważmy, że pierwszy człon w funkcji zysku (4.3.19) nie zależy od wielkości zamówienia. Jak już wspomnieliśmy w sytuacji zdefiniowanej w problemie 4.3.2, przedsiębiorstwo sprzedaje swoim klientom tyle energii, ile wynosi zapotrzebowanie na nią – niezależnie od tego, ile jej zamówi wcześniej. Przychód z tej sprzedaży nie zależy więc od wielkości zamówienia z i z punktu widzenia problemu wyboru optymalnej wartości z^* ten człon funkcji zysku (4.3.19) jest stały i nie ma wpływu na podejmowaną decyzję. W związku z tym w dalszej części bieżącego punktu będziemy go pomijać. Pozbędziemy się także minusa przed całym pozostałym wyrażeniem, zamieniając funkcję zysku $R(y, z)$ na funkcję kosztu $C(y, z) = - R(y, z)$:

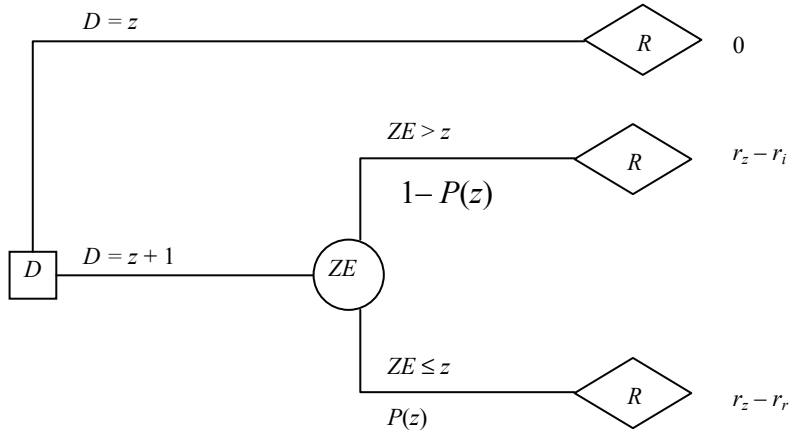
$$C(y,z) = r_z z - \max(z - y,0) r_r + \max(y - z,0) r_i \tag{4.3.20}$$

Podobnie jak w przypadku problemu 4.3.1, zamówienia dokonujemy, nie znając jeszcze faktycznej wielkości zapotrzebowania na energię elektryczną przez naszych odbiorców. Zakładamy naturalnie, że znamy jego prognozowany rozkład prawdopodobieństwa, określony za pomocą funkcji gęstości rozkładu $p(y)$ lub dystrybuanty $P(y)$, który może zostać wykorzystany do oceny ryzyka skutków decyzji i wyboru optymalnej wielkości zamówienia. Naszym kryterium w tym przypadku będzie więc wartość oczekiwana kosztów zamówienia energii, określona przy użyciu następującej zależności, analogicznej do (4.3.2) z poprzedniego punktu:

$$E(C(ZE,z)) = \int_0^\infty \left(r_z z - \max(z - y,0) r_r + \max(y - z,0) r_i\right) p(y)\, dy \tag{4.3.21}$$

Oczywiście całka w (4.3.21) obliczana jest od 0 do ∞, ponieważ wielkość zamówienia z musi być liczbą nieujemną.

Podsumowując więc obecne rozważania, należy stwierdzić, że aby rozwiązać zagadnienie decyzyjne sformułowane w problemie 4.3.2, musimy znaleźć taką wielkość zamówienia zakupu energii z^*, $z^* \geq 0$, dla której spodziewana wartość kosztów (4.3.21) będzie jak najmniejsza. Optymalna wielkość zamówienia ma, podobnie jak w problemie 4.3.1, minimalizować ryzyko popytowe wynikające z niepewności prognozowanego zapotrzebowania na energię elektryczną.



Rysunek 4.3.3. Drzewo decyzyjne dla analizy krańcowej w przypadku zadania określania optymalnej wielkości zamówienia w problemie 4.3.2
Źródło: opracowanie własne

Z jednej strony, określając wolumen zamówienia, należy brać pod uwagę ryzyko przeszacowania popytu i zakontraktowania zbyt dużych ilości energii, która nie zostanie sprzedana odbiorcom i będzie odsprzedana ze stratą (drugi człon w wyrażeniu podcałkowym w (4.3.21)). Z drugiej strony, trzeba rozważyć ryzyko niedoszacowania prognozy zapotrzebowania, co skutkuje podwyższonymi kosztami zakupu energii z innych źródeł (trzeci człon w wyrażeniu podcałkowym w (4.3.21)).

Znalezienie rozwiązania przedstawionego zagadnienia optymalizacji stochastycznej w warunkach ryzyka, podobnie jak w przypadku problemu rozważanego w punkcie poprzednim, nie jest zadaniem trudnym i również da się je rozwiązać na wiele sposobów. Można zastosować metody oparte na bezpośredniej minimalizacji spodziewanych kosztów (4.3.21). Wymagałyby one raczej wcześniejszego przekształcenia (4.3.21) do postaci odpowiadającej (4.3.9) z poprzedniego punktu. My jednak również i w tym przypadku zastosujemy metodę opartą na analizie krańcowej (Bartkiewicz 1999a; Bartkiewicz, Czajkowska i inni 2004).

W przypadku rozważanego obecnie problemu 4.3.2 mówimy, rzecz jasna, o analizie kosztów krańcowych, to znaczy zastosowanie tego podejścia polega na zbadaniu, jak zmieniają się spodziewane koszty przy zmianie wielkości zamówienia zakupu energii elektrycznej o jedną jednostkę. Rozważmy więc, podobnie jak w poprzednim przykładzie, dwie alternatywy decyzyjne, $D = z$ oraz $D = z + 1$. Przyjmując, że pierwszy wariant daje koszt odniesienia równy 0, zastanówmy się, co się zmieni w przypadku zamówienia dodatkowej jednostki energii.

Powyższa analiza dla sytuacji decyzyjnej w problemie 4.3.2 została przedstawiona w formie drzewa decyzyjnego na rysunku 4.3.3. Oczywiście tak samo jak w poprzednim punkcie, jeśli za punkt odniesienia weźmiemy koszt otrzymywany dla wielkości zamówienia zakupu energii z, to zwiększenie zamówienia o 1 daje dwie różne możliwości dotyczące zmiany kosztu – w zależności od faktycznego poziomu zapotrzebowania na energię ZE (Bartkiewicz, Gontar, Matusiak, Pamuła, Zieliński 2004; Bartkiewicz, Gontar, Matusiak, Zieliński 2001a).

Jeżeli zapotrzebowanie na energię okaże się mniejsze od wielkości zamówienia z bądź równe mu (dolna gałąź w węźle ZE), to dodatkowa zakupiona jednostka energii nie będzie mogła zostać sprzedana naszym odbiorcom. Zgodnie z warunkami problemu 4.3.2, możemy odsprzedać ją z pewną stratą i w ten sposób odzyskać część kosztów z nią związanych, uzyskując przychód z tej sprzedaży, który wynosi na każdej jednostce energii r_r złotych. Kupujemy więc dodatkowo jedną jednostkę energii po koszcie r_z, a sprzedajemy ją z przychodem r_r, czyli zmiana kosztów wynosi $r_z - r_r$. Pamiętajmy przy tym, że wartość ta będzie dodatnia, ponieważ jednostkowy przychód ze sprzedaży nadwyżki jest niższy od normalnego kosztu jednostkowego zakupu, tj. $r_r < r_z$. Prawdopodobieństwo takiego wyniku równe jest $\Pr(ZE \le z)$, czyli dystrybuancie spodziewanego rozkładu zapotrzebowania na energię, $P(z)$.

Jeżeli zapotrzebowanie na energię będzie większe od wielkości zamówienia z (górna gałąź w węźle ZE), to znaczy, że zapotrzebowanie to musi wynosić co najmniej $z + 1$, czyli dodatkowa jednostka energii nie tylko może, ale musi zostać dostarczona odbiorcom. Gdybyśmy nie zwiększyli wolumenu zamówienia, to musielibyśmy kupić tę jednostkę z kosztem r_i. Kupujemy zatem dodatkowo jedną jednostkę energii po koszcie r_z, a nie musimy jej kupować z kosztem r_i. Zmiana kosztów wynosi $r_z - r_i$. Zauważmy, że tym razem zmiana ta będzie ujemna, ponieważ zgodnie z warunkami problemu 4.3.2, koszt jednostkowy zakupu dodatkowej energii jest wyższy od normalnego kosztu jednostkowego zakupu, tj. $r_i > r_z$. Prawdopodobieństwo takiego stanu rzeczy wynosi $1 - \Pr(ZE \le z)$, czyli $1 - P(z)$.

Wyznaczmy wartość oczekiwaną kosztów dla zwiększonego zamówienia $z + 1$. W naszym przypadku, analizując sytuację przedstawioną na drzewie decyzyjnym na rysunku 4.3.3, możemy zapisać:

$$E(C(ZE, z+1)) = (r_z - r_i)(1 - P(z)) + (r_z - r_r)P(z) \tag{4.3.22}$$

Analogicznie jak w poprzednim podrozdziale i tym razem nietrudno zauważyć, że optymalna wielkość zamówionej energii elektrycznej z^*, minimalizująca wartość oczekiwaną kosztów zakupu energii dla tego zamówienia $E(C(ZE, z))$ (określoną przez zależność (4.3.21)), w warunkach rozkładu ryzyka popytowego

$P(z)$ będzie równa wartości z^*, która stanowi rozwiązanie równania różnicowego:

$$E(C(ZE, z+1)) - E(C(ZE, z)) = 0 \tag{4.3.23}$$

Interpretacja (4.3.22) jest dosyć oczywista i wynika niemal bezpośrednio z analizy naszego drzewa na rysunku 4.3.3. Rozumowanie przebiega analogicznie do przedstawionego w poprzednim problemie 4.3.1. Przeanalizujmy ponownie decyzję $D = z + 1$ na rysunku 4.3.3. Widzimy, że w górnej gałęzi węzła *ZE*, w której zapotrzebowanie na energię będzie większe od wielkości zamówienia z, zwiększenie wielkości zamówienia energii z o jedną jednostkę daje pewien ujemny przyrost kosztów w wysokości $r_z - r_i$. Natomiast w dolnej gałęzi węzła *ZE*, gdy zapotrzebowanie na energię okazuje się mniejsze od wielkości zamówienia z lub mu równe, przyrost kosztów jest dodatni i wynosi $r_z - r_r$. W sumie ważonej (4.3.22), określającej oczekiwany przyrost kosztów przy zwiększeniu zamówienia o 1, mnożymy więc pewną wartość ujemną przez prawdopodobieństwo $\Pr(ZE > z)$ oraz pewną wartość dodatnią przez prawdopodobieństwo $\Pr(ZE \le z)$.

Jeżeli wielkość zamówienia z jest mała, dużo niższa od spodziewanego zapotrzebowania na energię, to oczywiście prawdopodobieństwo niedoszacowania zamówienia i znalezienia się poniżej poziomu zapotrzebowania, $\Pr(ZE > z)$, jest duże, zaś prawdopodobieństwo przeszacowania i przekroczenia popytu, $\Pr(ZE \le z)$ – niewielkie. Co za tym idzie, w (4.3.22) mnożymy dodatni człon przez bardzo małe prawdopodobieństwo, bliskie wartości 0, zaś ujemny – przez duże, bliskie 1. Zmiana wartości oczekiwanej kosztów dla zwiększonej wielkości zamówienia $E(C(ZE, z+1))$ będzie więc ujemna.

Ponieważ przyrost wartości oczekiwanej kosztów przy wzroście wielkości zamówienia jest ujemny, opłaca się zamówienie zwiększyć. Pamiętajmy jednak, że wraz ze wzrostem wielkości zamówienia z maleje prawdopodobieństwo sprzedaży jeszcze jednej, dodatkowej jednostki energii, a rośnie prawdopodobieństwo, że nie zostanie ona sprzedana. Przyrosty spodziewanych kosztów $E(C(ZE, z+1)) - E(C(ZE, z))$ będą nadal ujemne, ale coraz większe, ponieważ w zależności (4.3.22) maleje prawdopodobieństwo $\Pr(ZE > z)$ członu ujemnego, natomiast rośnie prawdopodobieństwo $\Pr(ZE \le z)$ członu dodatniego. Dla pewnej wartości z zmiana ta stanie się w końcu większa od zera. W tej sytuacji dalsze zwiększanie wielkości zamówienia energii powodowałoby wzrost wartości oczekiwanej kosztów. W związku z tym nie powinniśmy tego robić.

Podsumowując – widzimy, że zwiększanie wielkości zamówienia opłaca się, kiedy przyrost wartości oczekiwanej kosztów jest ujemny, zaś przestaje się opłacać z chwilą, gdy stanie się on dodatni. Optymalna wielkość zamówienia z^* wyznaczana jest wobec tego przez punkt równowagi, to znaczy taką wartość z, dla której przyrost spodziewanych kosztów jest równy 0, czyli dla rozwiązania

równania (4.3.23). Sama zatem logika problemu 4.3.2 i kształt funkcji kosztów (4.3.21) gwarantują, że minimum kosztów istnieje oraz rozwiązanie równania (4.3.23) wyznacza jego wartość.

Równanie (4.3.23), podobnie jak w poprzednim problemie, ma charakter liniowy względem $P(z)$ i jego rozwiązanie również stanowi dosyć proste zadanie. Pamiętając, że wartość oczekiwana kosztów dla wielkości zamówienia energii równej z jest w naszym przypadku poziomem odniesienia kosztów równym 0, tzn. $E(C(ZE, z)) = 0$. Podstawiając $E(C(ZE, z+1))$ z zależności (4.3.22), otrzymujemy:

$$(r_z - r_i)(1 - P(z)) + (r_z - r_r)P(z) = 0 \tag{4.3.24}$$

Wykonując proste przekształcenia algebraiczne, wyznaczamy rozwiązanie (4.3.24), względem $P(z)$:

$$\begin{aligned} &(r_z - r_i) - P(z)(r_z - r_i) + (r_z - r_r)P(z) = 0 \\ &r_z - r_i - P(z)r_z + P(z)r_i + P(z)r_z - P(z)r_r = 0 \\ &r_z - r_i + P(z)r_i - P(z)r_r = 0 \\ &P(z)r_i - P(z)r_r = r_i - r_z \\ &P(z)(r_i - r_r) = r_i - r_z \end{aligned} \tag{4.3.25}$$

skąd ostatecznie otrzymujemy zależność określającą z^*, czyli optymalną wielkość zamówienia energii elektrycznej dla problemu 4.3.2, która minimalizuje funkcję oczekiwanych kosztów jej zakupu (4.3.21):

$$P(z^*) = \frac{r_i - r_z}{r_i - r_r} \tag{4.3.26}$$

Pamiętajmy, że zgodnie z warunkami zdefiniowanymi w problemie 4.3.2, obowiązuje następujące uporządkowanie kosztów zakupu energii $r_r < r_z < r_i$. Jak więc widzimy, w sytuacji zdefiniowanej w tym problemie optymalna wielkość zamówienia energii elektrycznej z^* określana jest przez (4.3.26), to znaczy wyznaczona zostaje przez wartość dystrybuanty $P(z)$ przewidywanego rozkładu zapotrzebowania na energię, odpowiadającą stosunkowi różnicy między kosztem energii dodatkowej r_i a kosztem energii normalnej r_z do rozmiarów całego przedziału kosztów zakupu, jaki w ogóle rozważamy $r_z - r_r$.

Przypomnijmy, że w podrozdziale 3.4 wskazywaliśmy, iż w przypadku modeli neuronowych i neuronowo-rozmytych krótkoterminowej prognozy zapotrzebowania na energię istnieją przesłanki, by przyjąć normalny rozkład zapotrzebowania. Fakt ten wymaga oczywiście weryfikacji w odniesieniu do konkretnego modelowanego systemu elektroenergetycznego, ale nasze doświadczenia wskazują na to, że na ogół nie ma z tym większych problemów.

Podobnie jak w poprzednim punkcie, dla rozkładu normalnego prognozy zapotrzebowania zagadnienie określenia optymalnej wielkości zamówienia energii elektrycznej, w warunkach określonych w problemie 4.3.2, wymaga jedynie obliczenia prostej poprawki dotyczącej prognozy otrzymywanej na wyjściu modelu. Dystrybuanta rozkładu normalnego jest funkcją różnowartościową, więc możemy rozwiązać równanie (4.3.26) względem z, wyznaczając wartość jej funkcji odwrotnej. Przekształcając (4.3.26), otrzymujemy optymalną wartość z^* (Matusiak, Bartkiewicz 2001; Bartkiewicz, Gontar, Matusiak, Zieliński 2002; Bartkiewicz, Matusiak 2003; Bartkiewicz, Gontar, Matusiak, Pamuła, Zieliński 2004; Bartkiewicz, Matusiak 2004):

$$z^* = P^{-1}\left(\frac{r_i - r_z}{r_i - r_r}\right) \tag{4.3.27}$$

gdzie P^{-1} stanowi funkcję odwrotną do dystrybuanty, kwantyl, prognozowanego rozkładu prawdopodobieństwa zapotrzebowania na energię elektryczną, $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$. Wartość oczekiwana rozkładu szacowana jest przez wyjście modelu prognostycznego (prognozę) $f(\mathbf{x}, \mathbf{w})$, zaś $\sigma_y(\mathbf{x})$ – odchyleniem standardowym prognozy. Dla dosyć dużej grupy typów sieci neuronowych i neuronowo-rozmytych, jako predyktorów krótkoterminowego zapotrzebowania na energię, metody szacowania odchylenia standardowego wyjścia modelu analizowaliśmy i badaliśmy w rozdziale 3.

Normalizując rozkład prawdopodobieństwa zapotrzebowania na energię $N(f(\mathbf{x}, \mathbf{w}), \sigma_y(\mathbf{x}))$, otrzymany na podstawie analizy działania modelu prognostycznego, do standardowego rozkładu normalnego $N(0, 1)$, możemy zapisać zależność (4.3.27) w równoważnej postaci. Optymalną wielkość zamówienia energii elektrycznej z^* wyznaczamy wówczas jako prostą korektę do prognozy otrzymanej z modelu neuronowego lub neuronowo-rozmytego:

$$z^* = f(\mathbf{x}, \mathbf{w}) + Q_{N(0,1)}\left(\frac{r_i - r_z}{r_i - r_r}\right) \cdot \sigma_y(\mathbf{x}) \tag{4.3.28}$$

gdzie $Q_{N(0,1)}(\cdot)$ jest kwantylem rozkładu normalnego standardowego $N(0, 1)$, natomiast $f(\mathbf{x}, \mathbf{w})$ i $\sigma_y(\mathbf{x})$, tak samo jak w przypadku (4.3.27), są wyjściem sieci neuronowej (neuronowo-rozmytej) i jego odchyleniem standardowym, wyznaczonym za pomocą metod opisanych w rozdziale 3.

Ponownie zwróćmy uwagę na konsekwencje zależności (4.3.26) dla zagadnienia wykorzystania prognoz krótkoterminowego zapotrzebowania na energię w procesie planowania wielkości zamówienia na rynku energii. Sytuacja jest tu zbliżona do rozważanej w poprzednim punkcie podczas analizy zależności (4.3.14). Jak już mówiliśmy, modele prognostyczne krótkoterminowego zapotrzebowania na energię budowane są na podstawie kryteriów minimalizacji błędu statystycznego (w większości przypadków kwadratowego) prognozy. Wyjście modelu prognostycznego $f(\mathbf{x}, \mathbf{w})$ możemy więc interpretować (patrz rozdział 3.1.1) jako przybliżenie wartości oczekiwanej $E(ZE / \mathbf{x})$ w przewidywanym rozkładzie prawdopodobieństwa zapotrzebowania na energię ZE, przy określonym $\mathbf{x}$, wzorcu wartości zmiennych wejściowych modelu.

Tymczasem, jeżeli przyjrzymy się zależności (4.3.26), to od razu widzimy, że również w obecnym przypadku, w warunkach zdefiniowanych w problemie 4.3.2, określenie strategii wyboru wielkości zamówienia z^* wyłącznie w odniesieniu do prognozy wartości oczekiwanej zapotrzebowania na energię $E(ZE / \mathbf{x})$ może mieć charakter nieoptymalny. Prognoza otrzymywana z modelu $f(\mathbf{x}, \mathbf{w})$ będzie odpowiadać optymalnej wielkości zamówienia z^* jedynie w przypadku, w którym stosunek kosztów jednostkowych zakupu energii, występujący po prawej stronie zależności (4.3.26), czyli $(r_i - r_z) / (r_i - r_r)$, jest równy 0,5. Ponadto rozkład prawdopodobieństwa zapotrzebowania musi mieć charakter symetryczny, w przeciwnym razie optymalną wielkość zamówienia energii wyznaczać będzie mediana, a nie wartość oczekiwana rozkładu.

Jak nietrudno zauważyć warunek $(r_i - r_z) / (r_i - r_r) = 0{,}5$ odpowiada sytuacji, gdzie $r_i - r_z = r_z - r_r$, tj. różnica między kosztem jednostkowym dodatkowego i normalnego zakupu energii elektrycznej jest taka sama jak kwota jednostkowa, której nie uda się odzyskać, sprzedając ewentualny nadmiar zamówionej energii. W tej sytuacji koszty zarówno niedoszacowania, jak i przeszacowania wolumenu zamówienia energii rozkładają się symetrycznie, w związku z tym optymalną wielkość zamówienia stanowi najbardziej prawdopodobna wartość zapotrzebowania na energię odbiorców, czyli $E(ZE / \mathbf{x})$ albo $f(\mathbf{x}, \mathbf{w})$.

Jeżeli zaś $(r_i - r_z) / (r_i - r_r)$, czyli stosunek kosztów jednostkowych zakupu energii występujący po lewej stronie zależności (4.3.26), jest mniejszy od 0,5, to dla każdej zamówionej jednostki energii elektrycznej różnica między kosztem dodatkowego i normalnego zakupu jest mniejsza od kwoty, której nie uda się odzyskać, sprzedając ewentualny nadmiar w ostatniej chwili, to znaczy $r_i - r_z < r_z - r_r$. Przyjęcie jako wielkości zamówienia wartości oczekiwanej (prognozy) zapotrzebowania na energię skutkowałoby jednakowym traktowaniem odchyleń dodatnich i ujemnych zamówienia od rzeczywistego popytu. W takim przypadku z jednakowym prawdopodobieństwem narażamy się na ryzyko przeszacowania bądź niedoszacowania zamówienia.

W rozważanej sytuacji natomiast zamówienie zbyt małej ilości energii powoduje podwyższenie spodziewanych kosztów zakupu tylko o różnicę między

kosztem jednostkowym dodatkowego i normalnego zakupu $r_i - r_z$ na każdej jednostce. Z kolei w przypadku przeszacowania wielkości zamówienia o takim samym rozmiarze w kategoriach bezwzględnych ryzykujemy kwotą, której nie uda się odzyskać, sprzedając ewentualny nadmiar zamówionej energii, czyli $r_z - r_r$ na każdej nadmiarowej jednostce, co stanowi wartość większą.

Podobnie jak w poprzednim punkcie, w problemie 4.3.1 mamy więc do czynienia z nierównomiernymi kosztami finansowymi takiego samego odchylenia prognozy na plus i na minus, czego w modelach statystycznych, które prognozują wartość oczekiwaną zapotrzebowania na energię elektryczną w ogóle nie bierze się pod uwagę. Gdy $(r_i - r_z) / (r_i - r_r)$ jest mniejsze od 0,5, optymalna wielkość zamówienia z^*, maksymalizująca oczekiwane zyski, aby skompensować wyższe koszty przeszacowania prognozy, musi być mniejsza od przewidywanej przez model prognostyczny wartości oczekiwanej zapotrzebowania na energię. Być może częściej zamówienie okaże się za małe i zmusi nas to do dokupienia energii, ale za to rzadziej zamówienie będzie zbyt duże, co skutkuje koniecznością sprzedaży nadwyżki, a to generuje wyższe koszty. Formuła (4.3.26) określa właśnie optymalną wartość (w świetle niepewności posiadanej prognozy) tej obniżki.

Z drugiej jednak strony, gdy stosunek kosztów zakupu energii zapisany po lewej stronie równania (4.3.26), $(r_i - r_z) / (r_i - r_r)$, wynosi powyżej 0,5, to różnica między kosztem jednostkowym dodatkowego i normalnego zakupu energii elektrycznej jest większa od kwoty jednostkowej, której nie uda się odzyskać, sprzedając ewentualny nadmiar zamówionej energii, to znaczy $r_i - r_z > r_z - r_r$. Tym razem niedoszacowanie zamówienia spowoduje wyższy wzrost kosztów zakupu energii dodatkowej niż przeszacowanie i sprzedaż posiadanej nadwyżki ze stratą. Z tego powodu, rzecz jasna, optymalna wielkość zamówienia z^* powinna być wyższa od prognozowanej przez model wartości oczekiwanej rozkładu prawdopodobieństwa zapotrzebowania, $E(ZE / \mathbf{x})$.

Podsumowując nasze rozważania, możemy powiedzieć, że decyzja odnośnie do wielkości zakupu energii elektrycznej w warunkach ryzyka popytowego w istotnym stopniu zależy od wartości współczynnika spodziewanych przez decydenta kosztów zakupu energii, występującego po lewej stronie równania (4.3.26). Jeżeli stosunek kosztów $(r_i - r_z) / (r_i - r_r)$ jest równy 0,5, optymalna strategia polega na określeniu wielkości zakupu na poziomie wartości oczekiwanej zapotrzebowania $E(ZE / \mathbf{x})$, czyli prognozy otrzymywanej na wyjściu modelu $f(\mathbf{x}, \mathbf{w})$. Jeżeli natomiast współczynnik ten odchyla się w kierunku wartości 0 lub 1, pojawia się nierównowaga kosztów finansowych zakupu spowodowanych niedoszacowaniem bądź przeszacowaniem prognozy i powinniśmy zastosować formułę (4.3.26) w celu określenia, odpowiednio, niższej bądź wyższej wielkości zamówienia. Ponieważ w przypadku prognoz zapotrzebowania na energię otrzymanych za pomocą sieci neuronowych i neuronowo-rozmytych często mamy do czynienia z rozkładem normalnym błędu prognozy, więc równanie

(4.3.26) możemy wówczas rozwiązać, korzystając z zależności (4.3.28) (lub alternatywnie (4.3.27)).

Zaprezentowane w naszej rozprawie podejście do rozwiązania zadania określania optymalnej wielkości zakupu w warunkach sformułowanych w problemie 4.3.2 (oczywiście również w problemie 4.3.1 w poprzednim punkcie) nie jest jedynym możliwym. Cała kwestia wynika z faktu, że w powszechnie stosowanych w procesie tworzenia modeli krótkoterminowej prognozy zapotrzebowania na energię kwadratowych funkcjach kosztu jednakowo traktuje się dodatnie i ujemne błędy prognozy, co niekoniecznie musi się przenosić na podobną równowagę kosztów zakupu energii elektrycznej. Rozwiązaniem alternatywnym może być więc zastosowanie w fazie treningu sieci neuronowej lub neuronowo-rozmytej funkcji kosztu uwzględniającej warunki finansowe problemu, tak by sama prognoza miała charakter optymalny pod względem kosztów.

W niniejszej książce nie przedstawiamy szerzej tego podejścia. W Katedrze Informatyki Uniwersytetu Łódzkiego prowadzone były jednak przez doktor Bożenę Matusiak prace również i w tym kierunku (Matusiak, Bartkiewicz 2001; Bartkiewicz, Gontar, Matusiak, Zieliński 2002; Bartkiewicz, Matusiak 2003; Bartkiewicz, Matusiak 2004). Osiągane efekty okazały się porównywalne z zastosowaniem korekty prognozy przy użyciu (4.3.28). Należy tutaj jednak zwrócić uwagę na kilka problemów związanych z odejściem od standardowego treningu sieci (lub, ogólniej, modelu prognostycznego) przy kwadratowej funkcji błędu:

– finansowe funkcje kosztu dla danego problemu decyzyjnego nie mają charakteru standardowego i nie ma gotowych algorytmów treningowych dla sieci neuronowych (neuronowo-rozmytych), które mogłyby zostać zastosowane w ich przypadkach,

– tego rodzaju funkcje kosztu zazwyczaj są trudne w zastosowaniu; mogą być w pewnych punktach nieróżniczkowalne, a nawet nieciągłe (patrz dla przykładu (4.3.1) czy też (4.3.20)); może to utrudniać zastosowanie w procesie uczenia modelu gradientowych metod optymalizacji; we wzmiankowanych pracach zastosowano układ hybrydowy, w którym do treningu sieci neuronowej wykorzystano algorytm genetyczny; należy jednak wspomnieć, że wykorzystanie poprawki (4.3.28) również wymaga wykonania dodatkowych obliczeń, opisanych w rozdziale 3, potrzebnych do oszacowania odchylenia standardowego wyjścia modelu prognostycznego $\sigma_y(\mathbf{x})$,

– podejście oparte na analizie niepewności modelu i użyciu poprawki (4.3.28) jest bardziej elastyczne; model prognostyczny podlega uczeniu standardową i uniwersalną metodą, podobny charakter mają algorytmy szacowania rozkładu; uzyskany system może być wykorzystywany do różnych decyzji podejmowanych w różnych warunkach; model prognostyczny tworzony z wykorzystaniem funkcji kosztu wynikających z danego problemu decyzyjnego może być stosowany wyłącznie do tego problemu.

Jak więc widzimy, zastosowanie równania (4.3.26) lub jego rozwiązania, określonego przez (4.3.28) (lub alternatywnie (4.3.27)), pozwala oszacować wielkość zamówienia, z którą związane będą niższe koszty zakupu niż w przypadku bezpośredniego wykorzystania prognozy wartości oczekiwanej zapotrzebowania na energię. Powstaje tu pytanie, jakie możemy uzyskać oszczędności, stosując zaproponowaną strategię. Oczywiste wydaje się, że będą one zależały od wartości współczynnika kosztów zakupu energii występującego po lewej stronie równania (4.3.26). Gdy stosunek kosztów $(r_i - r_z) / (r_i - r_r)$ jest równy 0,5, wynik (4.3.26) pokrywa się ze strategią polegającą na przyjęciu wielkości zakupu na poziomie wartości oczekiwanej zapotrzebowania $E(ZE / \mathbf{x})$, czyli prognozy otrzymywanej na wyjściu modelu $f(\mathbf{x}, \mathbf{w})$ (zakładamy, że rozkład prognozy jest symetryczny i jego wartość oczekiwana pokrywa się z medianą). Im bardziej odchyla się on w stronę wartości 0 lub 1, tym bardziej opłacalne powinno być zastosowanie (4.3.26).

Aby sprawdzić zakres możliwych oszczędności, przeprowadzono symulacje wielkości kosztów niezbilansowania jednej ze spółek dystrybucyjnych w przypadku zastosowania (Bartkiewicz, Gontar, Matusiak, Zieliński i inni 2002; Bartkiewicz, Matusiak 2003; Bartkiewicz, Matusiak 2004):

– ANN – strategii zakupu energii bezpośrednio na podstawie prognoz zapotrzebowania otrzymanych z wykorzystaniem standardowej warstwowej sieci perceptronowej,

– GANN – strategii zakupu energii na podstawie prognoz zapotrzebowania uzyskanych za pomocą sieci perceptronowej uczonej z wykorzystaniem algorytmu genetycznego z funkcją kosztu (4.3.20),

– ANN_ADJ – strategii zakupu energii na podstawie prognoz zapotrzebowania uzyskanych za pomocą sieci perceptronowej, skorygowanych przy użyciu (4.3.28); oszacowanie odchylenia standardowego prognozy $\sigma_y(\mathbf{x})$ otrzymane zostało za pomocą zależności (3.3.24), to jest z zastosowaniem metody delta z dokładnym wyznaczeniem hesjanu błędu (patrz punkt 3.3.2) oraz modelu czynnika losowego o zmiennym odchyleniu standardowym modelowanym przez dodatkową sieć neuronową (patrz punkt 3.4.3).

Symulacje wykonano dla trzech scenariuszy związanych z różnymi wartościami współczynnika kosztów zakupu energii $(r_i - r_z) / (r_i - r_r)$. Pierwszy z nich ($C_I$), dla wartości współczynnika 0,82, oznacza znacznie wyższe koszty niedoszacowania niż przeszacowania zamówienia. W drugim scenariuszu (C_{II}) przyjmujemy wartość współczynnika wynoszącą 0,24, tzn. znacznie wyższe koszty przeszacowania niż niedoszacowania zamówienia. Trzeci scenariusz (C_{III}) dla wartości 0,5 oznacza sytuację równowagi.

Tabela 4.3.1. Koszty niezbilansowania zakupu energii elektrycznej w problemie 4.3.2 dla różnych wariantów współczynnika $(r_i - r_z) / (r_i - r_r)$, dla wybranych godzin (w PLN)

Scenariusz	Strategia	Koszty dla godziny				
		6.00	12.00	17.00	18.00	24.00
$C_I = 0{,}82$	ANN	12 224	21 461	24 341	19 827	11 409
	GANN	9 998	16 996	21 774	18 309	10 991
	ANN_ADJ	9 029	14 007	15 080	15 252	11 213
$C_{II} = 0{,}24$	ANN	17 302	26 359	28 262	30 230	20 791
	GANN	14 909	23 476	28 615	28 476	14 584
	ANN_ADJ	14 034	21 736	24 900	23 409	13 183
$C_{III} = 0{,}5$	ANN	21 167	34 340	37 799	35 856	23 020
	GANN	22 363	37 959	48 121	44 495	24 623
	ANN_ADJ	21 170	34 328	37 775	35 869	23 062

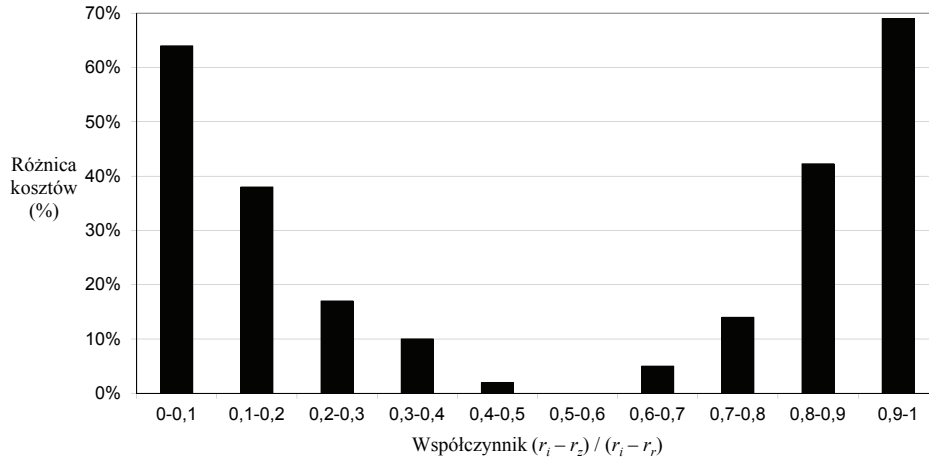
Źródło: opracowanie własne na podstawie W. Bartkiewicz, B. Matusiak, *Short-term load forecasting for energy markets*, [w:] L. Rutkowski, J. Kacprzyk (eds), *Neural Networks and Soft Computing*, Berlin–Heidelberg 2003, s. 790–795.

Wyniki symulacji dla kilku wybranych godzin doby przedstawione zostały w tabeli 4.3.1. Jak widzimy, przy rzeczywistym modelu rozkładu prawdopodobieństwa zapotrzebowania na energię elektryczną wyniki w scenariuszach C_I oraz C_{II}, to jest w warunkach znacznej nierównowagi kosztów niedoszacowania lub przeszacowania zakupu, wskazują na możliwość dosyć wyraźnej poprawy efektywności kosztowej zakupów. Porównanie efektów strategii opartej wyłącznie na prognozie ANN oraz strategii wykorzystującej korektę (4.3.28), wynikającą z niepewności prognozy ANN_ADJ, wskazuje, że przy wartościach współczynnika kosztów zakupu energii $(r_i - r_z) / (r_i - r_r)$ odchylających się od poziomu 0,5, to drugie podejście pozwala, potencjalnie, na redukcję kosztów niezbilansowania zamówienia z rzeczywistym popytem rzędu nawet do 20–30%. Strategia oparta na budowie modelu prognostycznego bezpośrednio z finansową funkcją kosztów (GANN) dała w większości przypadków zbliżone efekty do skorygowanej wielkości zakupu ANN_ADJ.

Zauważmy, że w scenariuszu C_{III} symulacje efektów niezbilansowania dla strategii ANN, GANN oraz ANN_ADJ są bardzo zbliżone. Jest to zgodne z wcześniejszymi oczekiwaniami, ponieważ w scenariuszu tym wykorzystuje się wartość współczynnika kosztów zakupu energii $(r_i - r_z) / (r_i - r_r)$ równą 0,5, co odpowiada równowadze kosztów niedoszacowania lub przeszacowania wielkości zakupu.

Jak więc widzimy, przy dużych lub małych wartościach współczynnika $(r_i - r_z) / (r_i - r_r)$, niedoszacowanie lub przeszacowanie zakupu w stosunku do faktycznego zapotrzebowania odbiorców może generować znaczne koszty tego

niezbilansowania. Strategia określania wielkości zakupu ANN_ADJ, polegająca na zastosowaniu korekty wynikającej z niepewności prognozy (zależność (4.3.28)), oferuje możliwość znacznej redukcji tych dodatkowych kosztów. Aby dokładniej przyjrzeć się potencjalnym korzyściom z prezentowanej metody, przeprowadziliśmy kolejną symulację kosztów niezbilansowania dla metod ANN_ADJ i ANN. Ponownie dla rzeczywistej prognozy zapotrzebowania na energię jednej ze spółek dystrybucyjnych określiliśmy wartość stosunku $(r_i - r_z) / (r_i - r_r)$, koszty niezbilansowania przy strategii, w której wykorzystuje się bezpośrednio prognozę ANN, oraz po wykorzystaniu korekty (4.3.28) (ANN_ADJ). Następnie przeanalizowaliśmy średni stosunek kosztów niezbilansowania zakupu energii ANN/ANN_ADJ przy różnych poziomach (przedziałach) wartości współczynnika $(r_i - r_z) / (r_i - r_r)$.



Rysunek 4.3.4. Redukcja kosztów niezbilansowania zakupu energii w problemie 4.3.2 dla ANN_ADJ w stosunku do ANN przy różnych przedziałach wartości współczynnika $(r_i - r_z) / (r_i - r_r)$

Źródło: opracowanie własne na podstawie B. Matusiak, W. Bartkiewicz, *Linking neural predictors with decision models: The short-term load forecasting case study*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk (eds), *Present-Day Problems of Power Engineering (APE '2001)*, vol. 3, Gdańsk–Jurata 2001, s. 109–116

Wyniki przedstawione zostały na rysunku 4.3.4. Widzimy, że nawet niewielkie odchylenie współczynnika $(r_i - r_z) / (r_i - r_r)$ od wartości 0,5 potencjalnie może dać kilkuprocentową różnicę kosztów na korzyść metody ANN_ADJ. Większe odchylenia mogą skutkować nawet kilkudziesięcioprocentową różnicą.

Zwróćmy jednak uwagę, że mówimy tutaj o korzyściach potencjalnych. Ich pełne osiągnięcie wymaga dokładnej znajomości cen r_z, r_i, r_r. W pewnych sytuacjach zresztą jest to nawet możliwe. W przypadku jednak gdy chodzi o transakcje na giełdzie energii lub rynku bilansującym, możemy mówić raczej

o znajomości prognoz i oszacowań tychże wielkości. Może to obniżać nieco uzyskiwane korzyści, jednakże przy dobrej jakości przewidywań cen nadal podejście, w którym wykorzystuje się korektę prognozy zapotrzebowania (4.3.28), czyli ANN_ADJ, powinno dawać wyraźnie lepsze wyniki.

4.3.3. Optymalna alokacja zakupionej energii na większą liczbę niepewnych popytów

W poprzednich punktach bieżącego podrozdziału mówiliśmy o określaniu wielkości zakupu (zamówienia) towaru o krótkiej trwałości i przy braku możliwości magazynowania – a tak jest w przypadku energii elektrycznej – w warunkach ryzyka popytowego. Obecnie zajmiemy się nieco innym problemem. Jeżeli już zdecydujemy się na pewną wielkość zamówienia, to rodzi się pytanie, w jaki sposób rozdzielić je na określoną liczbę odbiorców o niepewnym (wyznaczonym przez obarczoną ryzykiem prognozę) zapotrzebowaniu.

Podobnie jak w poprzednich punktach bieżącego podrozdziału, problemy występujące w tego rodzaju zagadnieniach oraz sposób wykorzystania prognoz zapotrzebowania do ich rozwiązywania zaprezentujemy w postaci problemów, omawiając w każdym z nich wybrane charakterystyki analizowanej dziedziny. Rozpoczniemy również od wersji zagadnienia alokacji zamówienia na dwóch niezależnych odbiorców w warunkach niepewności prognoz (oszacowań) ich zapotrzebowania, przy założeniu, że odbiorcy ci płacą za energię zakupioną.

Problem 4.3.3

Przyjmijmy, że przedsiębiorstwo obrotu energią elektryczną zaplanowało zakup z jednostek energii w postaci krótkoterminowej transakcji z pewnego źródła na rynku w określonej godzinnej jednostce rozliczeniowej. Przedsiębiorstwo obsługuje dwa odbiory o niezależnych zapotrzebowaniach określonych przez zmienne losowe Y_1 oraz Y_2. Znane są prognozy popytu dla obu odbiorów w rozważanym okresie dostawy w postaci funkcji gęstości rozkładów prawdopodobieństwa zapotrzebowania na energię $p_1(y_1)$ oraz $p_2(y_2)$ lub ich dystrybuant $P_1(y_1)$ oraz $P_2(y_2)$. Zysk jednostkowy sprzedaży energii dotyczący poszczególnym odbiorów wynosi odpowiednio r_{s1} oraz r_{s2}, przy czym wartości te mogą być różne (np. z powodu różnych kontraktów na dostawę, różnych kosztów dostawy itp.). Nie zmniejszając ogólności problemu, możemy uporządkować rozważane odbiory zgodnie z niemalejącym zyskiem jednostkowym z dostawy. Przyjmijmy więc, że spełniony jest warunek $r_{s1} \le r_{s2}$.

Zadanie polega oczywiście na zaplanowaniu rozdziału zakontraktowanej ilości energii z na oba odbiory, czyli określeniu takich wartości z_1, z_2, że $z_1 + z_2 = z$ oraz $z_1,\ z_2 \ge 0$, dla których osiągany jest maksymalny zysk. W rozważanym problemie przyjmiemy również, że łączna wielkość dostępnej energii z jest stała

i niezależnie od warunków oraz faktycznego zapotrzebowania nie może ulec zmianie. Z tego powodu w procesie optymalizacji analizować będziemy wyłącznie zależności w obrębie sprzedaży. Koszty zakupu są stałe, zawsze płacimy za z jednostek energii, więc możemy je pominąć.

Zakładamy również, że po ustaleniu planowanych wielkości z_1, z_2 pozostają one stałe. Niezależnie od rzeczywistego zapotrzebowania dotyczącego poszczególnych odbiorów, przydziały te nie mogą być zmieniane, zaś odbiory rozliczane są za energię faktycznie zakupioną. Przypomina to warunki problemu 4.3.1. Jeżeli na potrzeby któregoś z odbiorów przydzielona zostanie zbyt duża ilość energii, przekraczająca jego zapotrzebowanie, przedsiębiorstwo poniesie straty z powodu niewykorzystania części energii z rozdzielanej puli z. Gdy zaplanowane z_1 lub z_2 będzie zbyt małe w stosunku do zapotrzebowania danego odbioru, zakładamy, że przedsiębiorstwo nie będzie w stanie dostarczyć odbiorcom całej potrzebnej energii, poniesie w związku z tym straty z powodu niewykorzystania istniejących szans dodatkowej sprzedaży i uzyskania większego zysku. Niemożliwe są również przesunięcia w chwili fizycznej dostawy między wcześniej ustalonymi wielkościami z_1 i z_2.

Chwilowo ograniczymy się do zadania optymalnego rozdziału zamówienia na dwa niezależne popyty. W tej sytuacji możemy naturalnie oznaczyć ilość planowanej energii dla drugiego odbioru $z_2 = z - z_1$ i rozważać zadanie optymalizacyjne jednej zmiennej z_1. Przy okazji pozbywamy się również warunku ograniczającego naszego zadania optymalizacyjnego $z_1 + z_2 = z$.

Dla określonych wartości poziomu zapotrzebowania na energię elektryczną obu odbiorów, $Y_1 = y_1$ i $Y_2 = y_2$, oraz dla ustalonej ilości energii przydzielonej na potrzeby pierwszego odbioru z_1 funkcję zysku w naszym zadaniu decyzyjnym sformułowanym w problemie 4.3.3 możemy zapisać następująco:

$$R(y_1, y_2, z_1) = r_{s1} \min(y_1, z_1) + r_{s2} \min(y_2, z - z_1) \tag{4.3.29}$$

Interpretacja zależności (4.3.29) jest dosyć zbliżona do funkcji zysku (4.3.1) w problemie 4.3.1. Jeżeli energia przydzielona na potrzeby każdego z odbiorów będzie mniejsza od jego zapotrzebowania, sprzedajemy mu jedynie tyle energii, ile mu przydzielono. Gdy przydział energii będzie zbyt duży, odbiorca nie wykorzysta jej i pobierze tylko tyle, ile wynosi jego zapotrzebowanie. Tak jak mówiliśmy, pomijamy koszty zakupu całości energii, ponieważ są one stałe i nie mają wpływu na jej rozdział.

Oczywiście, podobnie jak w poprzednich przypadkach, w bieżącym podrozdziale zamówienie rozdzielamy, nie znając jeszcze rzeczywistej wielkości zapotrzebowania na energię elektryczną rozważanych odbiorców. Tak jak założyliśmy, znamy jednak ich prognozowane rozkłady w postaci funkcji gęstości rozkładów prawdopodobieństwa zapotrzebowania na energię $p_1(y_1)$ i $p_2(y_2)$ lub ich dystrybuant $P_1(y_1)$ i $P_2(y_2)$. Naszym kryterium będzie oczywiście wartość

oczekiwana zysków z dostawy energii, określona za pomocą następującej zależności:

$$E(R(y_1, y_2, z_1)) = \int_0^\infty \int_0^\infty \left(r_{s1} \min(y_1, z_1) + r_{s2} \min(y_2, z - z_1)\right) p(y_1, y_2) dy_1 dy_2 \qquad (4.3.30)$$

Całki w (4.3.21) obliczane są w przedziałach od 0 do ∞, ponieważ ilości energii przydzielanej na potrzeby poszczególnych odbiorów muszą być liczbami nieujemnymi.

W problemie 4.3.1 zakładamy, że zapotrzebowania na energię Y_1, Y_2 odnośnie do poszczególnych odbiorów są niezależne, więc korzystając z warunku niezależności $p(y_1, y_2) = p_1(y_1)p_2(y_2)$, funkcję wartości oczekiwanej zysku możemy poważnie uprościć. I tak przekształcając (4.3.30), otrzymujemy:

$$E(R(y_1, y_2, z_1)) = \int_0^\infty \int_0^\infty \left(r_{s1} \min(y_1, z_1) + r_{s2} \min(y_2, z - z_1)\right) p(y_1) p(y_2) dy_1 dy_2 = \qquad (4.3.31)$$

$$= r_{s1} \int_0^\infty \min(y_1, z_1) p(y_1) dy_1 \int_0^\infty p(y_2) dy_2 + r_{s2} \int_0^\infty \min(y_2, z - z_1) p(y_2) dy_2 \int_0^\infty p(y_1) dy_1$$

Ostatecznie więc pamiętając, że całka z funkcji gęstości prawdopodobieństwa w całej przestrzeni probabilistycznej równa jest 1, funkcję wartości oczekiwanej zysku ze sprzedaży energii otrzymujemy w następującej postaci:

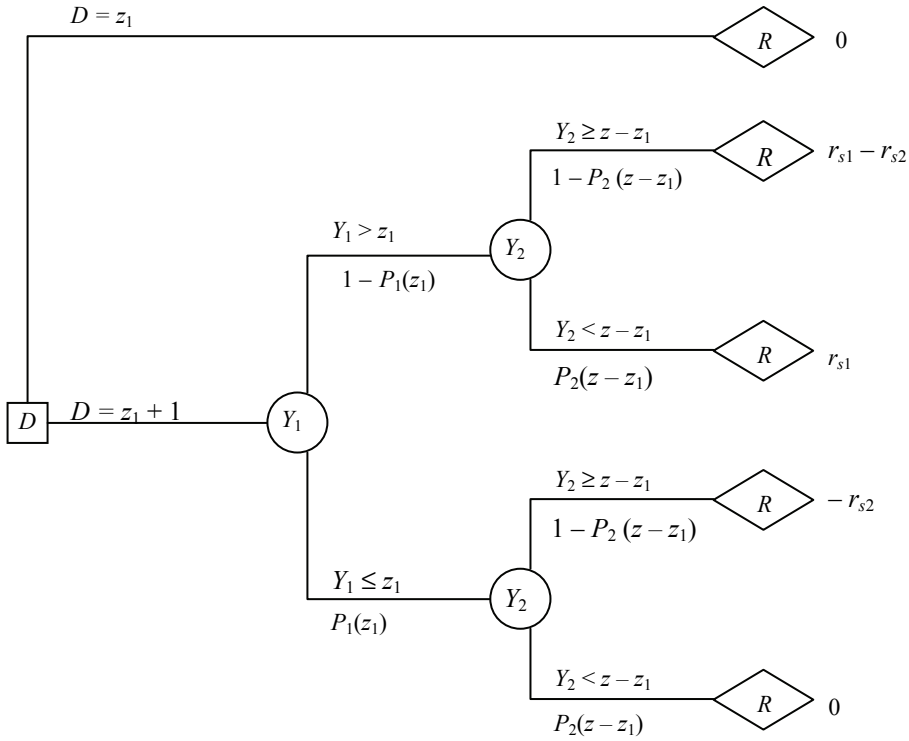
$$E(R(y_1, y_2, z_1)) = r_{s1} \int_0^\infty \min(y_1, z_1) p(y_1) dy_1 + r_{s2} \int_0^\infty \min(y_2, z - z_1) p(y_2) dy_2 \qquad (4.3.32)$$

Aby zatem rozwiązać zadanie decyzyjne sformułowane w problemie 4.3.3, musimy znaleźć z_1^*, $z_1^* \geq 0$, tzn. taką ilość planowanej energii przydzielonej odbiorowi pierwszemu, dla której spodziewana wartość zysków (4.3.32) będzie jak największa. Optymalny plan przydziału energii dla odbioru drugiego wyniesie wówczas $z_2^* = z - z_1^*$. Planowane wartości z_1^* i z_2^* będą określać najwyższą spodziewaną wartość zysków w świetle ryzyka związanego z niepewnością zapotrzebowania na energię Y_1, Y_2 dotyczącą poszczególnych odbiorów. Równoważą one ryzyko przeszacowania i niedoszacowania planowanego przydziału energii każdemu z odbiorców.

Podobnie jak w przypadku poprzednich problemów z bieżącego podrozdziału, zaprezentowane zadanie optymalizacji stochastycznej można rozwiązać na szereg sposobów. Również i tym razem zastosujemy metodę opartą na analizie

krańcowej, badając zmiany wartości oczekiwanej zysku przy jednostkowej zmianie wielkości planowanego przydziału energii z_1^* dla odbioru pierwszego. Zwróćmy jednak uwagę na pewną różnicę w stosunku do poprzednio rozważanych problemów. Analizując zmianę spodziewanych zysków, musimy tym razem uwzględnić wpływ dwóch zmiennych losowych Y_1, Y_2.

Rozważmy więc ponownie dwie alternatywy decyzyjne, $D = z_1$ oraz $D = z_1 + 1$. Przyjmując, że pierwszy wariant decyzji daje zysk odniesienia równy 0, zastanówmy się, co się zmieni w przypadku przeniesienia jednej jednostki energii między odbiorami.



Rysunek 4.3.5. Drzewo decyzyjne analizy krańcowej dla zadania optymalnej alokacji zamówienia na niezależne odbiory w problemie 4.3.3
Źródło: opracowanie własne

Analiza krańcowa dla sytuacji decyzyjnej występującej w problemie 4.3.3 przedstawiona została w formie drzewa decyzyjnego znajdującego się na rysunku 4.3.5. Jak widzimy, zwiększenie z_1 planowanego przydziału energii dla pierwszego odbioru o jedną jednostkę daje cztery różne możliwości, jeśli chodzi o zmianę zysku, w zależności od faktycznego poziomu zapotrzebowania na

energię dotyczącego obydwu odbiorów. Scharakteryzujemy je zgodnie z kolejnością od góry na rysunku 4.3.5. Pamiętajmy przy tym, że zwiększenie planowanego przydziału energii dla odbioru pierwszego powoduje zmniejszenie przydziału dla odbioru drugiego. Decyzja $D = z_1 + 1$ oznacza więc tylko przesunięcie jednostki energii z odbioru drugiego do pierwszego.

Rozważmy pierwszą gałąź drzewa od góry – dla rozwiązania $D = z_1 + 1$. Jeżeli zapotrzebowania na energię Y_1, Y_2 dla obydwu odbiorów będą większe od wyjściowych ilości energii z_1, $z - z_1$, przydzielonych im w planowanym rozdziale ($Y_1 > z_1$ oraz $Y_2 \geq z - z_1$), to decyzja o zwiększeniu o jeden przydziału z_1 dla pierwszego z nich spowoduje, że dodatkowa jednostka zostanie temu odbiorcy sprzedana, zwiększając nasz zysk o przychód jednostkowy r_{s1}. Zwróćmy jednak uwagę, że zapotrzebowanie drugiego odbioru jest na tyle duże, że tu również moglibyśmy sprzedać tę jednostkę, a więc z tego punktu widzenia, tracimy na tym przesunięciu – r_{s2}. Reasumując, zwiększenie z_1 o jedną jednostkę skutkuje zmianą zysku o kwotę $r_{s1} - r_{s2}$. Zwróćmy uwagę, że jest to wartość ujemna, ponieważ zgodnie z warunkami zdefiniowanymi w problemie 4.3.3, przychody jednostkowe ze sprzedaży energii dla drugiego odbioru są większe niż dla pierwszego, $r_{s1} < r_{s2}$.

Druga gałąź drzewa decyzyjnego dla tej alternatywy odpowiada sytuacji, w której zapotrzebowanie dotyczące pierwszego odbioru będzie na tyle duże, że kontrahent zakupi dodatkową planowaną dla niego jednostkę energii ($Y_1 > z_1$), natomiast zapotrzebowanie dotyczące odbioru drugiego będzie zbyt małe ($Y_2 < z - z_1$), więc odbiorca i tak nie wykorzysta rozważanej przez nas jednostki. W związku z tym przeniesienie tej jednostki z odbioru drugiego do pierwszego generuje zysk w kwocie przychodu jednostkowego r_{s1}.

Z kolei gałąź trzecia obrazuje sytuację, w której zapotrzebowanie pierwszego odbiorcy będzie za małe ($Y_1 \leq z_1$), wobec czego nie kupi on dodatkowej jednostki energii. Natomiast mogłaby ona zostać sprzedana odbiorcy drugiemu ($Y_2 \geq z - z_1$). W tym przypadku przeniesienie rozważanej jednostki energii spowoduje więc stratę równą kwocie przychodu jednostkowego drugiego odbioru $-r_{s2}$.

Ostatnia gałąź drzewa decyzyjnego dotyczy rozwiązania $D = z_1 + 1$. W tej gałęzi zapotrzebowania obu odbiorców są za małe, aby którykolwiek z nich kupił rozważaną jednostkę energii, mamy bowiem $Y_1 \leq z_1$, $Y_2 < z - z_1$. Przesunięcie tej jednostki pomiędzy odbiorami niczego więc nie zmienia, jeśli chodzi o wartość zysku. Jego zmiana wynosi więc w tym przypadku 0.

Podobnie jak w poprzednich przypadkach, wyznaczymy teraz wartość oczekiwaną zysku z decyzji o zwiększeniu przydziału energii dla pierwszego odbioru ($D = z_1 + 1$), ważąc poszczególne skutki tego wariantu decyzji prawdopodobieństwami ich realizacji. Ponieważ mamy tutaj do czynienia z dwiema zmiennymi losowymi Y_1, Y_2 określającymi zapotrzebowanie na energię poszczególnych odbiorców, musimy rozważyć je po kolei w poszczególnych gałęziach,

począwszy od liści drzewa, a następnie posuwając się wstecz, w kierunku korzenia. Ta charakterystyczna operacja na drzewach decyzyjnych określana jest często jako „zwijanie drzewa".

W naszym przypadku, rozważając sytuację przedstawioną na drzewie decyzyjnym na rysunku 4.3.3, najpierw przeanalizujemy wpływ zapotrzebowania na energię dotyczącego odbioru drugiego, wyznaczając wartość oczekiwaną decyzji w węzłach zmiennej Y_2. Dla górnej gałęzi, czyli gdy $Y_1 > z_1$, możemy więc zapisać:

$$E(R(Y_1 > z_1, y_2, z_1 + 1)) = (r_{s1} - r_{s2})(1 - P_2(z - z_1)) + r_{s1}P_2(z - z_1) \tag{4.3.33a}$$

Podobnie dla węzła zmiennej Y_2 w dolnej gałęzi (czyli gdy $Y_1 \leq z_1$) wartość oczekiwana decyzji wynosić będzie:

$$\begin{aligned} &E(R(Y_1 \leq z_1, y_2, z_1 + 1)) = \\ &= -r_{s2}(1 - P_2(z - z_1)) + 0 \cdot P_2(z - z_1) = -r_{s2}(1 - P_2(z - z_1)) \end{aligned} \tag{4.3.33b}$$

Wykorzystując teraz otrzymane wartości oczekiwane decyzji dla zmiennej Y_2 w poszczególnych gałęziach wychodzących z węzła Y_1, czyli zależności (4.3.33a) i (4.3.33b), możemy wyznaczyć spodziewane zyski w węźle zmiennej Y_1, a jednocześnie dla całej alternatywy decyzyjnej $D = z_1 + 1$.

$$\begin{aligned} &E(R(y_1, y_2, z_1 + 1)) = \\ &= E(R(Y_1 > z_1, y_2, z_1 + 1))(1 - P_1(z_1)) + E(R(Y_1 \leq z_1, y_2, z_1 + 1))P_1(z_1) = \\ &= \big((r_{s1} - r_{s2})(1 - P_2(z - z_1)) + r_{s1}P_2(z - z_1)\big)(1 - P_1(z_1)) - r_{s2}(1 - P_2(z - z_1))P_1(z_1) \end{aligned} \tag{4.3.34}$$

I znów, tak samo jak w poprzednich zagadnieniach dyskutowanych w bieżącym podrozdziale, aby rozwiązać zadanie decyzyjne sformułowane w problemie 4.3.3, czyli określić taką ilość planowanej energii przydzielonej pierwszemu odbiorowi z_1^*, dla której wartość oczekiwana zysku (4.3.32) będzie jak największa, musimy znaleźć wartość z_1^* stanowiącą rozwiązanie równania:

$$E(R(y_1, y_2, z_1 + 1)) - E(R(y_1, y_2, z_1)) = 0 \tag{4.3.35}$$

Pamiętając, że wartość oczekiwana zysku w alternatywie decyzyjnej o przydziale energii dla pierwszego odbioru w wysokości z_1 ($D = z_1$) jest poziomem referencyjnym równym 0, tzn. $E(R(y_1, y_2, z_1)) = 0$, oraz podstawiając

$E(R(y_1, y_2, z_1+1))$ z zależności (4.3.34), możemy zapisać równanie (4.3.35) w następujący sposób:

$$\big((r_{s1}-r_{s2})(1-P_2(z-z_1))+r_{s1}P_2(z-z_1)\big)(1-P_1(z_1))-r_{s2}(1-P_2(z-z_1))P_1(z_1)=0 \quad (4.3.36)$$

Uprośćmy (4.3.36) przy wykorzystaniu podstawowych przekształceń algebraicznych, grupując po jednej stronie równania człony zawierające dystrybuantę $P_1(z_1)$, po drugiej stronie natomiast $P_2(z - z_1)$.

$$\big((r_{s1}-r_{s2})(1-P_2(z-z_1))+r_{s1}P_2(z-z_1)\big)(1-P_1(z_1))-r_{s2}(1-P_2(z-z_1))P_1(z_1)=0$$

$$\big(r_{s1}-r_{s2}-r_{s1}P_2(z-z_1)+r_{s2}P_2(z-z_1)+r_{s1}P_2(z-z_1)\big)(1-P_1(z_1))+ \quad (4.3.37)$$

$$-r_{s2}P_1(z_1)+r_{s2}P_2(z-z_1)P_1(z_1)=0$$

$$r_{s1}-r_{s2}+r_{s2}P_2(z-z_1)-r_{s1}P_1(z_1)+r_{s2}P_1(z_1)-r_{s2}P_2(z-z_1)P_1(z_1)+$$

$$-r_{s2}P_1(z_1)+r_{s2}P_2(z-z_1)P_1(z_1)=0$$

$$r_{s1}-r_{s2}+r_{s2}P_2(z-z_1)-r_{s1}P_1(z_1)\ =0$$

$$r_{s1}-r_{s1}P_1(z_1)=r_{s2}-r_{s2}P_2(z-z_1)$$

Stąd otrzymujemy ostateczną postać zależności, która pozwala wyznaczyć optymalną wielkość planowanego przydziału energii elektrycznej dla pierwszego odbioru, maksymalizującą wartość oczekiwaną zysku $E(R(y_1, y_2, z_1))$ określoną wzorem (4.3.32). Optymalna wartość z_1^* musi stanowić rozwiązanie równania:

$$r_{s1}(1-P_1(z_1))=r_{s2}(1-P_2(z-z_1)) \quad (4.3.38)$$

Zwróćmy jednak uwagę na pewien aspekt otrzymanych przez nas wyników optymalnego rozwiązania problemu 4.3.3. Otóż równanie (4.3.38) niekoniecznie musi mieć rozwiązanie w interesującym nas przedziale [0, z], w którym zmienia się wartość przydziału energii dla pierwszego odbioru z_1. Pamiętajmy bowiem o tym, że funkcje $P_1(y_1)$ i $P_2(y_2)$ stanowią dystrybuanty rozkładu prawdopodobieństwa spełniające określone warunki. Są to funkcje niemalejące, przyjmujące wartości z przedziału [0, 1]. W związku z tym nakłada to również pewne warunki na zachowanie zależności występujących po obu stronach (4.3.38), w obrębie przedziału [0, z]:

– lewa strona (4.3.38) jest funkcją nierosnącą; dla lewego krańca przedziału, czyli $z_1 = 0$, jej wartość równa jest r_{s1}, a dla prawego krańca przedziału $z_1 = z$ opada ona do wartości $r_{s1}(1 - P_1(z))$,

– prawa strona (4.3.38) jest funkcją niemalejącą; dla $z_1 = 0$, czyli lewego krańca przedziału zmienności z_1, wynosi ona $r_{s2}(1 - P_2(z))$; odpowiednio, w prawym krańcu przedziału, dla $z_1 = z$, jej wartość wzrasta do r_{s2}.

Badając wartości zależności po obu stronach (4.3.38), na krańcach przedziału $[0, z]$, łatwo możemy otrzymać pełną strategię postępowania w celu otrzymania optymalnych pod względem spodziewanych zysków wartości rozdziału energii z_1^*, $z_2^* = 1 - z_1^*$ na oba odbiory w warunkach określonych w problemie 4.3.3:

– jeżeli $r_{s1} < r_{s2}(1 - P_2(z))$, to (4.3.38) nie ma rozwiązania, ponieważ z charakteru monotoniczności funkcji znajdujących się po obu stronach tego równania wynika, że dla wszystkich z_1 z przedziału $[0, z]$ jego lewa strona jest mniejsza od prawej; optymalną strategią w tym przypadku będzie oczywiście przydział całości energii odbiorowi drugiemu, czyli $z_1^* = 0$, $z_2^* = z$,

– analogiczną sytuację mielibyśmy w przypadku, gdy $r_{s2} < r_{s1}(1 - P_1(z))$; wówczas z kolei (4.3.38) nie miałoby rozwiązania, ponieważ dla wszystkich z_1 z przedziału $[0, z]$ jego lewa strona byłaby większa od prawej; zauważmy jednak, że powyższy warunek jest niemożliwy, ponieważ $P_1(z)$ musi być wartością z przedziału $[0, 1]$, więc implikowałby on, że $r_{s2} < r_{s1}$; w sformułowaniu problemu 4.3.3 uporządkowaliśmy rozważane odbiory zgodnie z niemalejącym zyskiem jednostkowym dostawy, czyli zachodzi odwrotny warunek $r_{s1} \le r_{s2}$,

– gdy $r_{s2}(1 - P_2(z)) \le r_{s1} \le r_{s2}$, równanie (4.3.38) musi mieć rozwiązanie w rozważanym przedziale $[0, z]$ i optymalna wielkość alokacji energii dla poszczególnych odbiorów z_1^* musi stanowić jego rozwiązanie; wielkość przydziału dla odbioru drugiego z_2^* wynosi oczywiście $z_2^* = 1 - z_1^*$.

Jeżeli rozwiązanie (4.3.38) istnieje, to można je otrzymać, stosując odpowiednie metody numeryczne. Równanie to jest, co prawda, nieliniowe, ale nie charakteryzuje go duży stopień trudności i można je rozwiązać stosunkowo prostymi metodami numerycznego rozwiązywania równań, znajdującymi się w powszechnie stosowanych pakietach obliczeniowych (nawet typu dodatek typu Solver w arkuszach kalkulacyjnych). Gdy rozkład prawdopodobieństwa prognozowanego zapotrzebowania na energię elektryczną ma charakter rozkładu normalnego, czyli jego dystrybuanta jest różnowartościowa, rozwiązanie może być tylko jedno, ponadto mamy je otoczone (musi znajdować się w przedziale $[0, z]$), co pozwala zastosować w prosty sposób choćby metodę bisekcji.

W problemie 4.3.3 ograniczyliśmy się do alokacji zamówienia energii na dwa popyty. Spróbujmy teraz przyjrzeć się uogólnieniu tego problemu na większą ich liczbę.

Problem 4.3.4

Przyjmijmy, że mamy do czynienia z sytuacją podobną do określonej w problemie 4.3.3, to znaczy szukamy optymalnego planu alokacji zamówienia energii elektrycznej o łącznej wielkości z, na n odbiorów o niezależnych zapotrzebowaniach, określonych przez zmienne losowe Y_1, Y_2, …, Y_n. Znane są prognozy popytu dla wszystkich odbiorów w rozważanym okresie dostawy w postaci funkcji gęstości rozkładów prawdopodobieństwa zapotrzebowania na energię $p_1(y_1)$, $p_2(y_2)$, …, $p_n(y_n)$ lub ich dystrybuant $P_1(y_1)$, $P_2(y_2)$, …, $P_n(y_n)$. Zysk jednostkowy sprzedaży energii poszczególnym odbiorcom wynosi dla każdego z nich odpowiednio r_{s1}, r_{s2}, …, r_{sn}, przy czym, podobnie jak w poprzednim problemie, nie zmniejszając ogólności rozważań, możemy uporządkować te odbiory zgodnie z niemalejącym zyskiem jednostkowym dostawy. Przyjmijmy wobec tego, że spełniony jest warunek $r_{s1} \le r_{s2} \le \ldots \le r_{sn}$.

Nasze zadanie polega tym razem na zaplanowaniu rozdziału zakontraktowanej ilości energii z na wszystkie odbiory, czyli znalezieniu zestawu wartości z_1, z_2, …, z_n, takich że $z_1 + z_2 + \ldots + z_n = z$, oraz z_1, z_2, …, $z_n \ge 0$, dla którego osiągany jest maksymalny zysk.

Kolejne założenia przyjmujemy z problemu 4.3.3, to znaczy zakładamy, że łączna wielkość dostępnej energii z jest stała i niezależnie od warunków oraz faktycznego zapotrzebowania nie może ulec zmianie. Przyjmujemy również, że po ustaleniu planowanych wielkości z_1, z_2, …, z_n one również są stałe. Niezależnie od rzeczywistego zapotrzebowania dotyczącego poszczególnych odbiorów nie mogą być one zmieniane. Jeżeli któremuś odbiorcy przydzielona zostanie zbyt duża ilość energii, przekraczająca jego zapotrzebowanie, przedsiębiorstwo poniesie straty z powodu niewykorzystania części zamówionej i zapłaconej energii. Gdy zaplanowane z_k będzie zbyt małe w stosunku do zapotrzebowania k-tego odbioru, przedsiębiorstwo poniesie straty z powodu niewykorzystania istniejących szans dodatkowej sprzedaży i uzyskania większego zysku. Niemożliwe są również przesunięcia w chwili fizycznej dostawy między wcześniej ustalonymi różnymi wielkościami z_k i z_j.

Aby zapewnić spełnienie warunku ograniczającego $z_1 + z_2 + \ldots + z_n = z$, analogicznie jak w poprzednim problemie, zmienną z_n, to znaczy wielkość energii przydzielonej w planie alokacji ostatniemu odbiorowi, wyłączymy z procesu optymalizacji, określając jej wartość jako $z_n = z - z'$, gdzie z' jest sztuczną zmienną, wprowadzoną wyłącznie dla uproszczenia zapisu, równą sumie przydziałów energii dla pozostałych $n - 1$ odbiorów:

$$z' = \sum_{j=1}^{n-1} z_j \tag{4.3.39}$$

Dla określonych wartości poziomów zapotrzebowania na energię elektryczną wszystkich odbiorców, $Y_1 = y_1$, $Y_2 = y_2$, …, $Y_n = y_n$, oraz dla ustalonej ilości energii przydzielonej n–1 pierwszym odbiorcom, z_1, z_2, …, z_{n-1}, funkcję zysku w zadaniu decyzyjnym sformułowanym w problemie 4.3.4 możemy zapisać następująco:

$$R(y_1,...,y_n,z_1,...z_{n-1}) = \sum_{j=1}^{n-1} r_{sj} \min(y_j, z_j) + r_{sn} \min(y_n, z - z') \tag{4.3.40}$$

Interpretacja funkcji zysku (4.3.40) jest w zasadzie identyczna jak w przypadku dwóch odbiorów (zależność (4.3.29) w poprzednim problemie). Jeżeli energia przydzielona każdemu odbiorcy będzie mniejsza od jego zapotrzebowania, otrzymuje on jedynie tyle energii, ile mu przydzielono, natomiast jeżeli przydział energii będzie zbyt duży, wykorzysta on tylko tyle, ile wynosi jego zapotrzebowanie. Pomijamy koszty zakupu całości energii, ponieważ są one stałe i nie mają wpływu na optymalizację jej rozdziału.

I znów pamiętajmy, że alokacji energii dokonujemy, nie znając rzeczywistej wielkości zapotrzebowania rozważanych odbiorców, a jedynie jego prognozowane rozkłady w postaci funkcji gęstości rozkładów prawdopodobieństwa zapotrzebowania na energię $p_1(y_1)$, $p_2(y_2)$, …, $p_n(y_n)$ lub ich dystrybuant $P_1(y_1)$, $P_2(y_2)$, …, $P_n(y_n)$. Jako kryterium optymalizacji rozdziału przyjmiemy więc, tak jak w poprzednich przypadkach, wartość oczekiwaną zysków z dostawy energii określoną w następujący sposób:

$$E(R(y_1,...,y_n,z_1,...z_{n-1})) = \tag{4.3.41}$$
$$= \int_0^{\infty} \left(\sum_{j=1}^{n-1} r_{sj} \min(y_j, z_j) + r_{sn} \min(y_n, z - z') \right) p(y_1, y_2, ..., y_n) dy_1 dy_2 ... dy_n$$

Nie znamy oczywiście $p(y_1, y_2, …, y_n)$, to jest gęstości łącznego rozkładu prawdopodobieństwa zapotrzebowania poszczególnych odbiorców, ale wykorzystując założenie o ich niezależności, możemy (4.3.41) zapisać:

$$E(R(y_1,...,y_n,z_1,...z_{n-1})) = \tag{4.3.42}$$
$$= \int_0^{\infty} \left(\sum_{j=1}^{n-1} r_{sj} \min(y_j, z_j) + r_{sn} \min(y_n, z - z') \right) p_1(y_1) p_2(y_2),..., p_n(y_n) dy_1 dy_2 ... dy_n$$

co pozwala nam na sformułowanie ostatecznej postaci funkcji celu dla zadania optymalnego rozdziału energii określonego w problemie 4.3.4:

$$E(R(y_1,...,y_n,z_1,...z_{n-1})) = \\ = \sum_{j=1}^{n-1} r_{sj} \int_0^{\infty} \min(y_j, z_j) p_j(y_j) dy_j + r_{sn} \int_0^{\infty} \min(y_n, z - z') p_n(y_n) dy_n \tag{4.3.43}$$

gdzie, jak pamiętamy ze wzoru (4.3.39), z' jest sumą przydziałów energii dla pierwszych n–1 odbiorów.

Funkcja wartości oczekiwanej zysku (4.3.43) ma postać niezbyt wygodną do jej optymalizacji. Problemy mogą sprawiać występujące w niej operacje minimum. Da się jednak zapisać ją w bardziej wygodnej formie. Zwróćmy bowiem uwagę, że korzystając z zależności (4.3.6) (w punkcie 4.3.1), możemy poszczególne całki w (4.3.43) przekształcić w następujący sposób:

$$\int_0^{\infty} \min(y_j, z_j) p_j(y_j) dy_j = \int_0^{z_j} (1 - P_j(y_j)) dy_j \quad , j = 1,...,n-1 \\ \int_0^{\infty} \min(y_n, z - z') p_n(y_n) dy_n = \int_0^{z-z'} (1 - P_n(y_n)) dy_n \tag{4.3.44}$$

Korzystając więc z powyższych zależności, możemy zapisać funkcję spodziewanych zysków (4.3.43) w następującej postaci:

$$E(R(y_1,...,y_n,z_1,...z_{n-1})) = \\ = \sum_{j=1}^{n-1} r_{sj} \int_0^{z_j} (1 - P_j(y_j)) dy_j + r_{sn} \int_0^{z-z'} (1 - P_n(y_n)) dy_n \tag{4.3.45}$$

Aby zatem rozwiązać zadanie alokacji określonej wielkości energii z na n niezależnych popytów, sformułowane w problemie 4.3.4, musimy znaleźć nieujemne wartości z_1^*, z_2^*, …, z_{n-1}^*, czyli takie ilości planowanej energii przydzielonej pierwszym n–1 odbiorom, dla której spodziewana wartość zysków (4.3.45) (lub alternatywnie (4.3.43)) będzie jak największa. Optymalny plan przydziału energii dla ostatniego, n-tego odbioru będzie wówczas, oczywiście, wynosił $z_n^* = z - z'^*$.

Do znalezienia optymalnego rozdziału energii $z_1^*, z_2^*, \ldots, z_{n-1}^*$ w problemie 4.3.4 nie możemy wykorzystać analizy krańcowej. Metoda ta nie nadaje się do użycia w problemach wielowymiarowych. Musimy więc zastosować odpowiednie algorytmy optymalizacyjne. Zauważmy przy tym, że stosowane w procedurach maksymalizacyjnych gradienty funkcji celu naszego zadania względem zmiennych decyzyjnych $z_1, z_2, \ldots, z_{n-1}$ nie muszą być szacowane numerycznie, ponieważ dosyć łatwo możemy wyznaczyć ich postać analityczną.

Pochodne cząstkowe naszej funkcji oczekiwanych zysków względem z_k, $k = 1, \ldots, n-1$ możemy łatwo wyznaczyć, korzystając z jej postaci (4.3.45):

$$\begin{aligned}
&\frac{\partial E(R(y_1,\ldots,y_n,z_1,\ldots z_{n-1}))}{\partial z_k} = \\
&= \sum_{j=1}^{n-1} r_{sj} \frac{\partial}{\partial z_k}\left(\int_0^{z_i}(1-P_j(y_j))dy_j\right) + r_{sn}\frac{\partial}{\partial z_k}\left(\int_0^{z-z'}(1-P_n(y_n))dy_n\right) = \\
&= r_{sk}(1-P_k(z_k)) - \frac{\partial z'}{\partial z_k} r_{sn}(1-P_n(z-z')) = \\
&= r_{sk}(1-P_k(z_k)) - r_{sn}(1-P_n(z-z'))
\end{aligned} \tag{4.3.46}$$

Pamiętajmy bowiem, że dla $k = 1, \ldots, n-1$,

$$\frac{\partial z'}{\partial z_k} = \frac{\partial}{\partial z_k}\left(\sum_{j=1}^{n-1} z_j\right) = 1 \tag{4.3.46a}$$

Rozwiązaniem alternatywnym dla bezpośrednich gradientowych metod szukania ekstremum może być znalezienie maksimum spodziewanego zysku poprzez rozwiązanie układu równań przyrównujących pochodne cząstkowe funkcji celu do 0:

$$\frac{\partial E(R(y_1,\ldots,y_n,z_1,\ldots z_{n-1}))}{\partial z_k} = 0, \qquad k = 1,\ldots,n-1 \tag{4.3.47}$$

Wykorzystując (4.3.46), możemy teraz zapisać układ (4.3.47) w postaci zestawu znanych nam już równań stanowiących uogólnioną formę zależności dla przypadku dwóch odbiorów (4.3.38):

$$r_{sk}(1-P_k(z_k)) = r_{sn}(1-P_n(z-z')) \ , \quad k=1,...,n-1 \tag{4.3.48}$$

W dwóch ostatnich problemach, 4.3.3 oraz 4.3.4, zakładaliśmy, że przydziały energii dla poszczególnych odbiorów (popytów) nie mogą być zmieniane w przypadku ich niezbilansowania z rzeczywistym zapotrzebowaniem. Innymi słowy, ewentualne braki energii pokrywane są przez innych uczestników rynku. W kolejnym problemie spróbujemy przeanalizować zmiany w strategii formułowania planu alokacji energii w sytuacji, w której ewentualne koszty bilansowania całej zaplanowanej alokacji spoczywają na nas. Podobnie jak w poprzednim przypadku, ograniczymy się chwilowo do sytuacji rozdziału ustalonej wielkości energii na dwa niezależne popyty.

Problem 4.3.5

Podobnie jak w poprzednich problemach omawianych w bieżącym punkcie, interesuje nas zaplanowanie rozdziału puli z jednostek energii elektrycznej na dwa odbiory o niezależnych zapotrzebowaniach określonych przez zmienne losowe Y_1 oraz Y_2. Znane są oczywiście prognozy popytu w okresie dostawy dla obu odbiorów w postaci funkcji gęstości rozkładów prawdopodobieństwa zapotrzebowania na energię $p_1(y_1)$ oraz $p_2(y_2)$ lub ich dystrybuant $P_1(y_1)$ oraz $P_2(y_2)$.

Zadanie polega więc na określeniu takich wartości przydziałów energii z_1, z_2 poszczególnym odbiorcom, że $z_1 + z_2 = z$ oraz $z_1, z_2 \geq 0$, dla których osiągamy maksymalną wartość spodziewanych skutków decyzji. Zakładamy również, że po ustaleniu planowanych wielkości z_1, z_2 są one również stałe. Niezależnie od rzeczywistego zapotrzebowania poszczególnych odbiorców, nie mogą być one zmieniane. Niemożliwe są również przesunięcia w chwili fizycznej dostawy między wcześniej ustalonymi wielkościami z_1 i z_2. Ewentualne niezbilansowania każdego odbioru wywołują więc niezbilansowania całej alokowanej puli energii z, nawet jeżeli jeden z odbiorców ma nadwyżkę, a drugi – niedobór.

Podobnie jak w przypadku problemu 4.3.2, zakładamy, że niezbilansowania całej kwoty energii z rozliczane są przez transakcje zakupu przez nas brakującej energii z kosztem jednostkowym r_i złotych (gdy przydział zaplanowany energii będzie zbyt mały w stosunku do zapotrzebowania tego odbioru) lub sprzedaży z zyskiem jednostkowym r_r złotych (jeżeli przydzielona zostanie zbyt duża ilość energii, przekraczająca zapotrzebowanie danego odbioru). Przyjmujemy przy tym, że, podobnie jak w problemie 4.3.2, zachodzi zależność: $r_r < r_z < r_i$, gdzie r_z jest kosztem jednostkowym zakupu energii w wyjściowym zamówieniu z jednostek.

Tak samo jak w przypadku problemu 4.3.3, oznaczmy ilość planowanej energii dla odbioru drugiego przez $z_2 = z - z_1$, dzięki czemu zamienimy nasze zadanie optymalizacyjne na zadanie jednej zmiennej z_1, jak również nie będzie-

my musieli brać pod uwagę warunku ograniczającego naszego zadania optymalizacyjnego $z_1 + z_2 = z$.

Zauważmy, że w sformułowaniu problemu 4.3.5 pominęliśmy informacje o zyskach jednostkowych ze sprzedaży energii obydwu odbiorcom. W obecnej sytuacji informacje te po prostu stają się zbędne, ponieważ sprzedaż ta odpowiada dokładnie ich zapotrzebowaniu i nie zależy od planowanego rozdziału wstępnej puli z jednostek energii. W konsekwencji więc, analogicznie jak w problemie 4.3.2, zastosujemy formułę optymalizacji kosztowej. Ponadto w rozważanych warunkach nieistotna jest kwota kosztu zakupu energii r_z. Wielkość alokowanej całej puli, a co za tym idzie jej koszt zakupu, również są stałe, niezależnie od sposobu jej planowanego rozdziału na oba odbiory.

Celem naszej optymalizacji jest więc znalezienie takiego rozdziału energii na oba odbiory: z_1, $z - z_1$, który daje minimalną wartość kosztów wyłącznie ich niezbilansowania. Dla danych konkretnych wartości poziomu zapotrzebowania na energię elektryczną obu odbiorców, $Y_1 = y_1$ i $Y_2 = y_2$, oraz dla ustalonej ilości energii przydzielonej pierwszemu odbiorcy z_1, funkcję kosztu w zadaniu decyzyjnym w problemie 4.3.5 możemy zapisać następująco:

$$C(y_1, y_2, z_1) = \max(y_1 - z_1, 0) r_i - \max(z_1 - y_1, 0) r_r + \max(y_2 - z + z_1, 0) r_i - \max(z - z_1 - y_2, 0) r_r \tag{4.3.49}$$

Znaczenie poszczególnych członów składowych zależności (4.3.49) jest dosyć oczywiste. Jeżeli przydział energii dla któregoś z odbiorców okazuje się niedoszacowany, czyli jego zapotrzebowanie na energię jest wyższe od przydzielonej mu wielkości, brakującą energię trzeba będzie dokupić, z kosztem jednostkowym r_i (człon pierwszy i trzeci w (4.3.49)), zwiększając koszt całej alokacji. Natomiast gdy przydział dla któregoś z odbiorców jest przeszacowany, czyli występuje nadmiar przydzielonej mu energii ponad popyt, nadwyżkę można odsprzedać z zyskiem jednostkowym r_r, odzyskując część kosztów wydanych na nadmiarową energię.

Ponieważ podobnie jak w poprzednich przypadkach, w bieżącym podrozdziale alokacji dokonujemy, nie znając jeszcze rzeczywistego zapotrzebowania na energię elektryczną dla rozważanych odbiorów, musimy wyznaczyć wartość oczekiwaną kosztów przy ich prognozowanych rozkładach prawdopodobieństwa. Nasza funkcja celu określona więc będzie za pomocą następującej zależności:

$$E(C(y_1,y_2,z_1)) = \int_0^{\infty} (\max(y_1 - z_1,0)r_i - \max(z_1 - y_1,0)r_r + \max(y_2 - z + z_1,0)r_i - \max(z - z_1 - y_2,0)r_r)p(y_1,y_2)dy_1dy_2 \quad (4.3.50)$$

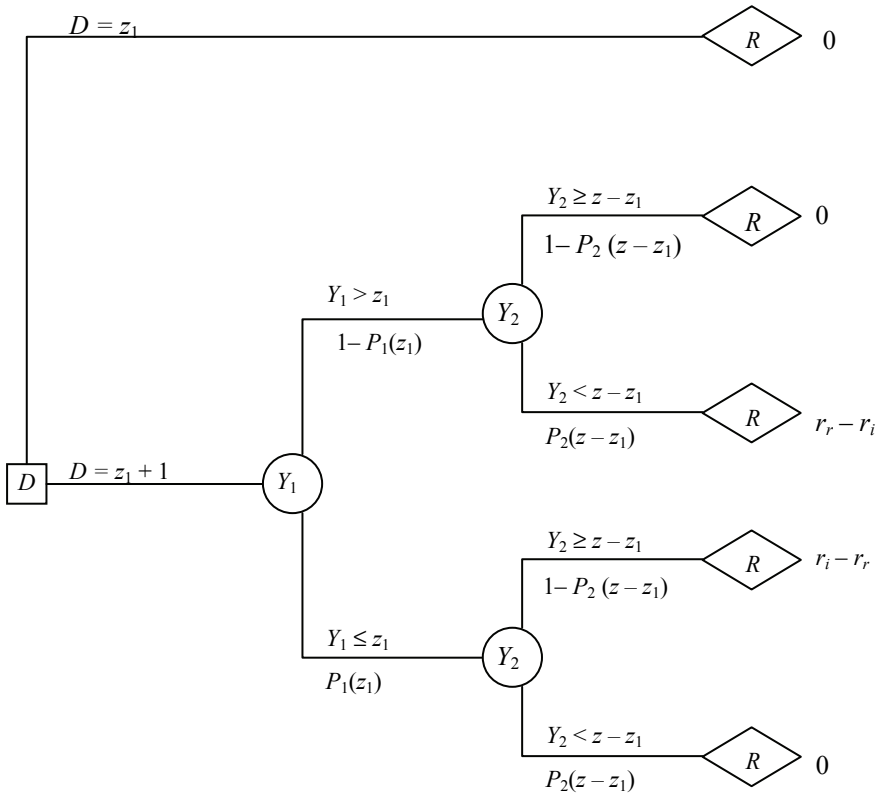
Korzystając z założenia o niezależności zmiennych losowych Y_1, Y_2, reprezentujących zapotrzebowanie na energię obu odbiorców, podobnie jak w poprzednich problemach, możemy (4.3.50) zapisać jako:

$$E(C(y_1,y_2,z_1)) = \int_0^{\infty} (\max(y_1 - z_1,0)r_i - \max(z_1 - y_1,0)r_r)p(y_1)dy_1 + \int_0^{\infty} (\max(y_2 - z + z_1,0)r_i - \max(z - z_1 - y_2,0)r_r)p(y_2)dy_2 \quad (4.3.51)$$

Podsumowując więc, aby znaleźć rozwiązanie zadania decyzyjnego określonego w problemie 4.3.5, musimy wyznaczyć taką wartość planowanego przydziału energii elektrycznej dla pierwszego z odbiorów z_1^*, $z_1^* \geq 0$, dla której spodziewana wartość kosztów (4.3.51) będzie jak najmniejsza. Oczywiście optymalna alokacja energii dla odbioru drugiego wyniesie wówczas $z_2^* = z - z_1^*$. Planowane wartości z_1^* i z_2^* będą określać najniższą spodziewaną wartość kosztów rozdziału energii na poszczególne odbiory w świetle ryzyka związanego z niepewnością zapotrzebowania na energię elektryczną Y_1, Y_2 dla tych odbiorów.

Jak już wielokrotnie wspominaliśmy, tego rodzaju zadanie optymalizacji stochastycznej można naturalnie rozwiązać na wiele sposobów. Również i w przypadku problemu 4.3.5 zastosujemy metodę opartą na analizie krańcowej, badając zmiany wartości oczekiwanej kosztów alokacji energii przy jednostkowej zmianie wielkości planowanego przydziału energii z_1^* dla odbioru pierwszego. Analizując zmianę spodziewanych kosztów, musimy uwzględnić wpływ dwóch zmiennych losowych Y_1, Y_2.

Drzewo decyzyjne dla analizy krańcowej odzwierciedlającej sytuację przedstawianą w problemie 4.3.5 znajduje się na rysunku 4.3.6. Jak zwykle, rozważmy na nim dwie możliwe alternatywy decyzyjne, $D = z_1$ oraz $D = z_1+1$. Przyjmując, że pierwszy wariant decyzji daje koszt odniesienia równy 0, zastanówmy się, jak zmienią się koszty rozdziału w przypadku przeniesienia jednej jednostki energii elektrycznej z odbioru drugiego do pierwszego.



Rysunek 4.3.6. Drzewo decyzyjne dla analizy krańcowej w przypadku problemu określania optymalnej alokacji zamówienia na niezależne dwa odbiory w problemie 4.3.5
Źródło: opracowanie własne

Podobnie jak w przypadku problemu 4.3.3, przesunięcie jednej jednostki w planowanym rozdziale energii pomiędzy obydwoma odbiorcami, w zależności od ich faktycznego poziomu zapotrzebowania, wywołuje cztery możliwe skutki, jeśli chodzi o zmianę kosztów. Omówimy je zgodnie z kolejnością od góry na rysunku 4.3.6.

Rozważmy więc pierwszą gałąź drzewa dla alternatywy decyzyjnej $D = z_1 + 1$. W tej gałęzi przesuwana jednostka energii mogłaby zostać sprzedana obydwu odbiorcom, ponieważ spełnione są warunki $Y_1 > z_1$ oraz $Y_2 \geq z - z_1$, czyli ich zapotrzebowanie jest większe od przydziałów. Dzięki decyzji o zwiększeniu przydziału dla pierwszego odbiorcy o jedną jednostkę ($D = z_1 + 1$) nie musimy więc zaspokajać zapotrzebowania tego odbioru zakupem bilansującym. Zmniejszamy łączny koszt alokacji energii o koszt zakupu dodatkowej jednostki r_i. Jednocześnie jednak musimy tę jednostkę zakupić, żeby zbilansować zapotrze-

bowanie drugiego odbioru, zwiększając koszt o kwotę r_i. Łącznie więc w tej gałęzi przesunięcie jednej jednostki energii daje w efekcie zmianę kosztu równą 0.

Druga gałąź drzewa decyzyjnego dla tego rozwiązania odpowiada sytuacji, w której $Y_1 > z_1$, natomiast $Y_2 < z - z_1$. Zapotrzebowanie pierwszego odbiorcy jest więc na tyle duże, że odbierze on dodatkową planowaną dla niego jednostkę energii, dzięki czemu nie musimy tu ponosić kosztu zakupu bilansującego. Koszt łączny maleje o kwotę r_i. Jednocześnie drugi odbiorca ma zbyt niskie zapotrzebowanie i w tej sytuacji przesuwana jednostka także nie zostałaby odebrana. Musielibyśmy sprzedać przesuwaną jednostkę w transakcji bilansującej. Ponieważ po przesunięciu do tej transakcji nie dojdzie, łączny koszt alokacji zwiększa się o kwotę r_r. W tej gałęzi łączna zmiana kosztów energii dla alternatywy decyzyjnej $D = z_1+1$ wynosić więc będzie $r_r - r_i$.

Gałąź trzecia dla alternatywy decyzyjnej $D = z_1+1$ obrazuje sytuację, w której $Y_1 \le z_1$ oraz $Y_2 \ge z - z_1$. Oznacza to, że zapotrzebowanie pierwszego odbiorcy będzie zbyt małe, wobec czego nie odbierze on dodatkowej jednostki energii. Zostanie ona sprzedana w transakcji bilansującej, co zmniejszy łączny koszt rozdziału o wartość r_r. Jednostkę tę zabraliśmy jednak odbiorcy drugiemu, którego zapotrzebowanie jest na tyle wysokie, że niezbędny będzie jej zakup w transakcji bilansującej. Łączny koszt rozdziału rośnie w związku z tym o kwotę r_i. Podsumowując, w tej gałęzi łączna zmiana kosztów energii dla alternatywy decyzyjnej $D = z_1+1$ wynosić będzie $r_i - r_r$.

Ostatnia gałąź drzewa decyzyjnego dla alternatywy decyzyjnej $D = z_1+1$ dotyczy sytuacji, w której $Y_1 \le z_1$ oraz $Y_2 < z - z_1$. W tej gałęzi zapotrzebowania obu odbiorców są więc za małe, aby którykolwiek z nich kupił rozważaną jednostkę energii. Niezależnie od tego, gdzie ją przydzielimy, będzie ona musiała być sprzedana w transakcji bilansującej. Sytuacja jest więc analogiczna do tej zdefiniowanej w gałęzi pierwszej. Zmiana kosztów związana z przesunięciem jednej jednostki energii pomiędzy pierwszym i drugim odbiorem również wynosi 0.

Jak zwykle, wyznaczymy teraz wartość oczekiwaną zysku z decyzji o przesunięciu jednostki energii dla pierwszego odbioru ($D = z_1 + 1$), ważąc poszczególne skutki tego wariantu decyzji prawdopodobieństwami ich realizacji. Podobnie jak w problemie 4.3.3, mamy tutaj do czynienia z dwiema zmiennymi losowymi Y_1, Y_2, określającymi zapotrzebowanie na energię poszczególnych odbiorców. Musimy więc rozważyć je po kolei w poszczególnych gałęziach, począwszy od liści drzewa na rysunku 4.3.6, a następnie posuwając obliczenia wstecz, w stronę jego korzenia.

Skutki naszej decyzji, czyli zmiany w kosztach całego rozdziału energii, w węzłach liści poddrzewa alternatywy decyzyjnej $D = z_1 + 1$ określone są zgodnie z przedstawionymi rozważaniami na rysunku 4.3.6. Jako następne w gałęziach poddrzewa występują węzły zmiennej Y_2 reprezentującej zapotrzebowanie na energię elektryczną odbioru drugiego. Rozpoczniemy w związku z tym od

przeanalizowania wpływu tej zmiennej, wyznaczając w jej węzłach wartość oczekiwaną skutków decyzji $D = z_1 + 1$. Dla węzła Y_2, znajdującego się w górnej gałęzi poddrzewa, czyli gdy $Y_1 > z_1$, spodziewana wartość kosztów wynosi:

$$E(C(Y_1 > z_1, y_2, z_1 + 1)) = (r_r - r_i)P_2(z - z_1) \tag{4.3.52a}$$

Analogicznie – dla węzła zmiennej Y_2 w dolnej gałęzi (czyli gdy $Y_1 \le z_1$) wartość oczekiwana skutków decyzji wynosić będzie:

$$E(C(Y_1 \le z_1, y_2, z_1 + 1)) = (r_i - r_r)(1 - P_2(z - z_1)) \tag{4.3.52b}$$

Posuwając się w kierunku korzenia w gałęzi poddrzewa alternatywy decyzyjnej $D = z_1 + 1$, widzimy następny węzeł – zapotrzebowania na energię dla odbioru pierwszego, czyli zmiennej losowej Y_1. Jak wynika z rysunku 4.3.6, w poddrzewie tym nie ma już innych węzłów. Wartość oczekiwana skutków decyzji w tym węźle określała więc będzie jednocześnie spodziewane koszty dla całej tej alternatywy decyzyjnej. Korzystając z zależności (4.3.52a) oraz (4.3.52b), możemy wyznaczyć tę wartość w następujący sposób:

$$\begin{aligned} &E(R(y_1, y_2, z_1 + 1)) = \\ &= E(R(Y_1 > z_1, y_2, z_1 + 1))(1 - P_1(z_1)) + E(R(Y_1 \le z_1, y_2, z_1 + 1))P_1(z_1) \\ &= (r_r - r_i)P_2(z - z_1)(1 - P_1(z_1)) + (r_i - r_r)(1 - P_2(z - z_1))P_1(z_1) \end{aligned} \tag{4.3.53}$$

Identycznie jak w poprzednich przypadkach, aby rozwiązać zadanie decyzyjne sformułowane w problemie 4.3.5, czyli znaleźć takie z_1, tj. ilość planowanej energii elektrycznej przydzielonej pierwszemu odbiorcy, dla której wartość oczekiwana kosztu (4.3.51) będzie jak najmniejsza, musimy znaleźć taką wartość z_1^*, która stanowi rozwiązanie równania:

$$E(R(y_1, y_2, z_1 + 1)) - E(R(y_1, y_2, z_1)) = 0 \tag{4.3.54}$$

Tradycyjnie pamiętając, że wartość oczekiwana zysku dla pierwszej alternatywy decyzyjnej $D = z_1$ jest poziomem odniesienia o wartości 0, czyli $E(R(y_1, y_2, z_1)) = 0$, oraz podstawiając $E(R(y_1, y_2, z_1+1))$ z zależności (4.3.53), możemy zapisać równanie (4.3.54) w następujący sposób:

$$(r_r - r_i)P_2(z - z_1)(1 - P_1(z_1)) + (r_i - r_r)(1 - P_2(z - z_1))P_1(z_1) = 0 \tag{4.3.55}$$

Równanie (4.3.55) możemy dalej uprościć przy wykorzystaniu podstawowych przekształceń algebraicznych, dążąc do pogrupowania po jednej jego stronie członów zależnych od dystrybuanty $P_1(z_1)$, po drugiej natomiast $P_2(z - z_1)$. Pamiętajmy przy tym, że zgodnie z założeniami sformułowanymi w problemie 4.3.5, zachodzi warunek $r_r < r_i$, a co za tym idzie $r_r \neq r_i$. W przeciwnym przypadku równanie (4.3.55) miałoby nieskończenie wiele rozwiązań.

$$\begin{aligned}
&(r_r - r_i)\big(P_2(z-z_1)(1-P_1(z_1)) - (1-P_2(z-z_1))P_1(z_1)\big) = 0 \\
&P_2(z-z_1)(1-P_1(z_1)) - (1-P_2(z-z_1))P_1(z_1) = 0 \\
&P_2(z-z_1) - P_2(z-z_1)P_1(z_1) - P_1(z_1) + P_2(z-z_1)P_1(z_1) = 0 \\
&P_2(z-z_1) - P_1(z_1) = 0
\end{aligned} \tag{4.3.56}$$

Stąd możemy określić ostateczną postać równania pozwalającego na wyznaczenie optymalnej wartości z_1, czyli planowanej alokacji energii elektrycznej dla pierwszego odbioru, dla której spodziewany koszt $E(C(y_1, y_2, z_1))$, określony wzorem (4.3.51), jest minimalny. Optymalny przydział z_1^* musi być rozwiązaniem równania:

$$P_1(z_1) = P_2(z - z_1) \tag{4.3.57}$$

Jak widzimy w problemie 4.3.5, optymalna wartość rozdziału energii na oba odbiory, z_1^*, $z - z_1^*$, nie zależy od kosztów transakcji rozliczających ich niezbilansowania. Jeżeli jednak bliżej zastanowimy się nad tą kwestią i spojrzymy na zmiany kosztów decyzji w drzewie na rysunku 4.3.6, to takie rozwiązanie nie powinno budzić większych wątpliwości. Pamiętajmy bowiem, że w problemie tym analizujemy alokację pewnej stałej kwoty energii z, zaś przesunięcia między odbiorami w obrębie tej kwoty, w warunkach sformułowanych w problemie 4.3.5, rodzą symetryczne zmiany w kosztach całego rozdziału. Optymalne rozwiązanie z_1^* zależy więc tylko od niepewności samych prognoz zapotrzebowania na energię odbiorców, opisywanych w równaniu (4.3.57) przez ich dystrybuanty rozkładu prawdopodobieństwa $P_1(y_1)$ i $P_2(y_2)$.

Równanie (4.3.57) wyznacza zatem optymalną wartość alokacji energii dla obu odbiorów, dającą najniższe spodziewane koszty rozdziału puli energii z w świetle ryzyka samej prognozy jako takiej. Planowane przydziały z_1^*, $z - z_1^*$ określane są na zasadzie jednakowego prawdopodobieństwa pokrycia zapotrzebowania poszczególnych odbiorów. Jeżeli np. oba odbiory mają dokładnie identyczną prognozę zapotrzebowania, to zgodnie z równaniem (4.3.57), zamówienie z należy rozdzielić między nie w równych częściach. Z właściwości

dystrybuanty rozkładu prawdopodobieństwa wynika, że równanie (4.3.57) musi mieć rozwiązanie, przy czym jeżeli dystrybuanty zapotrzebowania odbiorców są różnowartościowe (jak w przypadku rozkładu normalnego), rozwiązanie to jest jednoznaczne.

Na koniec, zastanówmy się jeszcze nad uogólnieniem problemu 4.3.5, alokacji puli z jednostek energii na dowolną skończoną liczbę n odbiorów, o niezależnych i niepewnych zapotrzebowaniach.

Problem 4.3.6

Szukamy więc optymalnego planu alokacji zamówienia energii elektrycznej o łącznej wielkości z na n odbiorów o niezależnych zapotrzebowaniach określonych przez zmienne losowe Y_1, Y_2, …, Y_n. Popyt dla wszystkich odbiorców w rozważanym okresie dostawy znany jest więc z dokładnością do prognozy, w postaci funkcji gęstości rozkładów prawdopodobieństwa zapotrzebowania na energię $p_1(y_1)$, $p_2(y_2)$, …, $p_n(y_n)$ lub ich dystrybuant $P_1(y_1)$, $P_2(y_2)$, …, $P_n(y_n)$.

Nasze zadanie polega więc oczywiście na zaplanowaniu rozdziału posiadanej ilości energii elektrycznej z na wszystkie odbiory, czyli znalezienie zestawu wartości $z_1, z_2, \ldots, z_n$ takich, że $z_1 + z_2 + \ldots + z_n = z$ oraz $z_k \geq 0$, $k = 0, \ldots, n-1$, dla którego osiągany jest minimalny koszt całego rozdziału.

Przyjmujemy również, że po ustaleniu planowanych wielkości $z_1, z_2, \ldots, z_n$ są one stałe. Niezależnie od rzeczywistego zapotrzebowania dotyczącego poszczególnych odbiorów nie mogą być one zmieniane. Niemożliwe są również przesunięcia w chwili fizycznej dostawy między wcześniej ustalonymi różnymi wielkościami z_k i z_j. Ewentualne niezbilansowanie któregokolwiek z odbiorów wywołuje więc niezbilansowanie całej alokowanej ilości energii z, nawet jeżeli jeden z odbiorców ma nadwyżkę, a inni niedobory.

Tak samo jak w przypadku problemu 4.3.5, zakładamy że niezbilansowania całej rozdzielanej ilości energii z rozliczane są przez transakcje zakupu brakującej energii z kosztem jednostkowym r_i (gdy zaplanowany przydział energii będzie zbyt mały w stosunku do zapotrzebowania tego odbioru) lub sprzedaży z zyskiem jednostkowym r_r złotych (jeżeli przydzielona zostanie zbyt duża ilość energii, przekraczająca zapotrzebowanie danego odbioru). Przyjmujemy przy tym, że, podobnie jak dla dwóch odbiorów, zachodzi zależność: $r_r < r_i$.

Również w przypadku problemu 4.3.6 przyjmiemy, że $z_n = z - z'$, gdzie z', tak samo jak w problemie 4.3.4, jest sztuczną zmienną, wprowadzoną wyłącznie dla uproszczenia zapisu, równą sumie przydziałów energii dla pozostałych $n - 1$ odbiorów:

$$z' = \sum_{j=1}^{n-1} z_j \tag{4.3.58}$$

Operacja ta pozwoli oczywiście na spełnienie warunku ograniczającego naszej decyzji, określonego przez równość $z_1 + z_2 + \ldots + z_n = z$ bez jego jawnego uwzględniania w sformułowaniu problemu.

Funkcja kosztu konkretnego rozdziału stanowi proste uogólnienie na przypadek wielowymiarowy funkcji kosztu z poprzedniego problemu 4.3.5, określonej za pomocą zależności (4.3.49). Również i w tej sytuacji pomijamy różne stałe elementy kosztów, niezależne od zmiennych decyzyjnych $z_1, z_2, \ldots, z_{n-1}$ i bierzemy pod uwagę wyłącznie koszt niezbilansowania poszczególnych odbiorów. Dla określonych wartości poziomów zapotrzebowania na energię elektryczną dotyczącego wszystkich odbiorów, $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, oraz dla ustalonego rozdziału energii przydzielonego $n - 1$ pierwszym odbiorom, $z_1, z_2, \ldots, z_{n-1}$, funkcję zysku w zadaniu decyzyjnym sformułowanym w problemie 4.3.6 możemy zapisać w następujący sposób:

$$C(y_1,...,y_n,z_1,...z_{n-1}) = \sum_{j=1}^{n-1}\left(\max(y_j - z_j,0)r_i - \max(z_j - y_j,0)r_r\right) + \max(y_n - z + z',0)r_i - \max(z - z' - y_n,0)r_r \tag{4.3.59}$$

Jak zwykle pamiętamy, rzecz jasna, że alokacji energii dokonujemy, nie znając rzeczywistej wielkości zapotrzebowania poszczególnych odbiorców, a jedynie jego prognozowane rozkłady, więc jako podstawę decyzji przyjmiemy wartość oczekiwaną kosztów rozdziału przy ich prognozowanych rozkładach prawdopodobieństwa, określonych przez funkcje gęstości rozkładu wyjścia modelu prognostycznego $p_1(y_1)$, $p_2(y_2)$, …, $p_n(y_n)$. Funkcja celu w naszym problemie decyzyjnym określona będzie zatem przy użyciu następującej zależności:

$$E(C(y_1,...,y_n,z_1,...z_{n-1})) = \int_0^{\infty}\left(\sum_{j=1}^{n-1}\left(\max(y_j - z_j,0)r_i - \max(z_j - y_j,0)r_r\right) + \max(y_n - z + z',0)r_i - \max(z - z' - y_n,0)r_r\right)p(y_1,y_2,...,y_n)dy_1dy_2...dy_n \tag{4.3.60}$$

Nie znamy gęstości łącznego rozkładu prawdopodobieństwa zapotrzebowania na energię wszystkich odbiorów, to jest $p(y_1, y_2, \ldots, y_n)$, ale wykorzystując założenie o ich niezależności, możemy rozłożyć łączny rozkład prawdopodobieństwa na iloczyn rozkładów brzegowych zapotrzebowania poszczególnych

odbiorów, $p(y_1, y_2, \ldots, y_n) = p_1(y_1) \cdot p_2(y_2) \cdot \ldots \cdot p_n(y_n)$, a następnie, korzystając z prostych właściwości całki, rozbić (4.3.60) na poszczególne człony składowe:

$$E(C(y_1,\ldots,y_n,z_1,\ldots z_{n-1})) =$$

$$= \sum_{j=1}^{n-1} \int_0^\infty \left(\max(y_j - z_j,0) r_i - \max(z_j - y_j,0) r_r\right) p_j(y_j) dy_j + \tag{4.3.61}$$

$$\int_0^\infty \left(\max(y_n - z + z',0) r_i - \max(z - z' - y_n,0) r_r\right) p_n(y_n) dy_n$$

Ponieważ do znalezienia minimum funkcji wartości oczekiwanej kosztu rozdziału energii elektrycznej na odbiory przydatne będzie wyznaczenie jej pochodnych, podobnie jak w problemie 4.3.4, doprowadźmy naszą funkcję celu (4.3.61) do nieco bardziej wygodnej postaci, pozbawionej występujących w niej operacji maksimum. Korzystając z definicji i podstawowych właściwości operacji maksimum i minimum, możemy zapisać następującą prostą zależność:

$$\max(t - c,0) = \max(t,c) - c = t + c - \min(t,c) - c = t - \min(t,c) \tag{4.3.62}$$

Wykorzystując otrzymaną zależność (4.3.62), możemy więc teraz przekształcić poszczególne człony w funkcji celu rozważanego obecnie zagadnienia decyzyjnego. Dla pierwszych n–1 składników zależności (4.3.61), czyli dla $j = 1, \ldots n$–1, możemy to zapisać w następujący sposób:

$$\int_0^\infty \left(\max(y_j - z_j,0) r_i - \max(z_j - y_j,0) r_r\right) p_j(y_j) dy_j = \tag{4.3.63a}$$

$$= \int_0^\infty \left(y_j r_i - \min(y_j, z_j) r_i - z_j r_r + \min(y_j, z_j) r_r\right) p_j(y_j) dy_j =$$

$$= \int_0^\infty \left(y_j r_i + \min(y_j, z_j)(r_r - r_i)\right) p_j(y_j) dy_j - z_j r_r$$

Analogicznie możemy przekształcić również ostatni składnik występujący w funkcji spodziewanych kosztów z rozdziału zamówienia:

$$\int_0^{\infty} \left(\max(y_n - z + z',0) r_i - \max(z - z' - y_n,0) r_r\right) p_n(y_n) dy_n =$$

$$= \int_0^{\infty} \left(y_n r_i + \min(y_n, z - z')(r_r - r_i)\right) p_n(y_n) dy_n - (z - z') r_r \tag{4.3.63b}$$

Podstawiając wyznaczone tu zależności (4.3.63) do funkcji celu naszego problemu (4.3.61), otrzymujemy:

$$E(C(y_1,...,y_n,z_1,...z_{n-1})) =$$

$$= \sum_{j=1}^{n-1} \left(\int_0^{\infty} \left(y_j r_i + \min(y_j, z_j)(r_r - r_i)\right) p_j(y_j) dy_j - z_j r_r \right) + \tag{4.3.64}$$

$$\int_0^{\infty} \left(y_n r_i + \min(y_n, z - z')(r_r - r_i)\right) p_n(y_n) dy_n - (z - z') r_r$$

Korzystając teraz z zależności (4.3.6) (patrz punkt 4.3.1), pozwalającej na obliczenie wartości oczekiwanej operacji minimum z wykorzystaniem dystrybuanty stosowanego rozkładu prawdopodobieństwa, otrzymujemy następującą postać funkcji spodziewanych kosztów dla problemu 4.3.6:

$$E(C(y_1,...,y_n,z_1,...z_{n-1})) =$$

$$= \sum_{j=1}^{n-1} \left(r_i \int_0^{\infty} y_j p_j(y_j) dy_j + (r_r - r_i) \int_0^{z_j} (1 - P_j(y_j)) dy_j - z_j r_r \right) + \tag{4.3.65}$$

$$r_i \int_0^{\infty} y_n p_n(y_n) dy_n + (r_r - r_i) \int_0^{z - z'} (1 - P_n(y_n)) dy_n - (z - z') r_r$$

co daje nam ostateczną postać funkcji wartości oczekiwanej kosztów rozdziału energii dla rozważanego problemu 4.3.6:

$$E(C(y_1,...,y_n,z_1,...z_{n-1})) =$$

$$= \sum_{j=1}^{n-1}\left(r_i E(Y_j) + (r_r - r_i)\int_0^{z_j}(1-P_j(y_j))dy_j - z_j r_r \right) + \tag{4.3.66}$$

$$r_i E(Y_n) + (r_r - r_i)\int_0^{z-z'}(1-P_n(y_n))dy_n - (z-z')r_r$$

Podsumowując nasze dotychczasowe rozważania dotyczące problemu 4.3.6 – aby rozwiązać zadanie alokacji określonej puli energii z na n niezależnych popytów w warunkach sformułowanych w tym problemie musimy znaleźć nieujemne wartości z_1^*, z_2^*, …, z_{n-1}^*, czyli takie kwoty planowanej energii przydzielonej pierwszym n–1 odbiorcom, dla której spodziewana wartość kosztów (4.3.61) (lub alternatywnie (4.3.66)) będzie jak najmniejsza. Optymalny plan przydziału energii dla ostatniego, n-tego odbiorcy będzie wówczas wynosił oczywiście $z_n^* = z - z'^*$.

Podobnie jak w przypadku problemu 4.3.4, również i obecnie do znalezienia optymalnego rozdziału energii z_1^*, z_2^*, …, z_{n-1}^* nie możemy zastosować analizy krańcowej. Jak już zwracaliśmy uwagę wcześniej, metoda ta nie nadaje się do wykorzystania w problemach wielowymiarowych. Do minimalizacji funkcji celu (4.3.61) lub alternatywnie (4.3.66) będziemy więc musieli zastosować odpowiednie algorytmy optymalizacyjne. Znów zauważmy, że wykorzystywane w procedurach minimalizacyjnych gradienty funkcji celu naszego zadania względem zmiennych decyzyjnych z_1, z_2, …, z_{n-1} nie muszą być szacowane numerycznie, ponieważ dosyć łatwo możemy wyznaczyć ich postać analityczną.

Pochodne cząstkowe naszej funkcji oczekiwanych kosztów względem z_k, $k = 1, \ldots, n-1$ możemy łatwo wyznaczyć, korzystając z jej postaci (4.3.66):

$$\frac{\partial E(R(y_1,...,y_n,z_1,...z_{n-1}))}{\partial z_k} =$$

$$= \sum_{j=1}^{n-1}\frac{\partial}{\partial z_k}\left(r_i E(Y_j) + (r_r - r_i)\int_0^{z_j}(1-P_j(y_j))dy_j - z_j r_r \right) + \tag{4.3.67}$$

$$\frac{\partial}{\partial z_k}\left(r_i E(Y_n) + (r_r - r_i)\int_0^{z-z'}(1-P_n(y_n))dy_n - (z-z')r_r \right)$$

Zauważmy, że w sumie znajdującej się w pierwszym członie tej zależności każdy składnik o indeksie j zależy wyłącznie od z_j. Dla wszystkich $j \neq k$ pochodne składników tej sumy względem z_k są więc równe 0. Ponadto ostatni człon (4.3.67) zależy od z_k tylko przez zmienną z'. Przypomnijmy bowiem (patrz zależność (4.3.58)), że zmienna z' stanowi sumę wszystkich z_j, $j = 1, \ldots, n-1$. Dla wszystkich $k = 1, \ldots, n-1$ pochodna z' względem z_k jest więc równa 1.

Z (4.3.67) otrzymujemy zatem dla $k = 1, \ldots, n-1$:

$$\frac{\partial E(R(y_1,...,y_n,z_1,...z_{n-1}))}{\partial z_k} =$$

$$= (r_r - r_i)(1 - P_k(z_k)) - r_r - (r_r - r_i)(1 - P_n(z - z')) + r_r = \tag{4.3.68}$$

$$= (r_r - r_i)\big((1 - P_k(z_k)) - (1 - P_n(z - z'))\big)$$

co daje ostateczną postać gradientu naszej funkcji wartości oczekiwanej kosztów rozdziału energii (4.3.66) w problemie 4.3.6:

$$\frac{\partial E(R(y_1,...,y_n,z_1,...z_{n-1}))}{\partial z_k} = (r_r - r_i)(P_n(z - z') - P_k(z_k)), k = 1,...,n-1 \tag{4.3.69}$$

Najwygodniejszym sposobem znalezienia optymalnego rozdziału energii z_1^*, z_2^*, …, z_{n-1}^* minimalizującego spodziewane koszty, czyli funkcję celu (4.3.61) lub alternatywnie (4.3.66), nie będzie chyba zastosowanie bezpośrednich metod optymalizacji przez przeszukiwanie gradientowe, lecz poprzez rozwiązanie układu równań przyrównujących pochodne cząstkowe funkcji celu do 0:

$$\frac{\partial E(R(y_1,...,y_n,z_1,...z_{n-1}))}{\partial z_k} = 0, \qquad k = 1,...,n-1 \tag{4.3.70}$$

Wykorzystując (4.3.69), możemy teraz zapisać równość (4.3.70) w postaci układu znanych nam już równań, stanowiących uogólnioną formę zależności dla przypadku dwóch odbiorów (4.3.57):

$$P_k(z_k) = P_n(z - z') \qquad , k = 1,...,n-1 \tag{4.3.71}$$

Optymalne wartości $z_1^*, z_2^*, \ldots, z_{n-1}^*$, minimalizujące koszt rozdziału energii w problemie 4.3.6, będą zatem rozwiązaniami układu (4.3.71). Oczywiście optymalny przydział dla ostatniego odbioru wynosi $z_n^* = z - z'^*$.

4.4. Podsumowanie

W rozdziale bieżącym omawiamy problem powiązań między prognozami a decyzjami podejmowanymi na ich podstawie. Nasz cel polegał na wykazaniu, że w zadaniach decyzyjnych znaczenie ma wykorzystanie nie tylko informacji o samej prognozie, ale również dodatkowej wiedzy na temat niepewności prognozy. Wiedza ta umożliwia oszacowanie związanego z niepewnością prognoz ryzyka podejmowanych decyzji oraz włączenie go w proces decyzyjny. Rozważania w tym rozdziale stanowią więc bezpośrednią kontynuację dyskusji z rozdziału trzeciego na temat modelowania niepewności predyktorów neuronowych i neuronowo-rozmytych. Potwierdziwszy w poprzednim rozdziale hipotezę badawczą dotyczącą przydatności przedstawionych metod do prognoz krótkoterminowego zapotrzebowania na energię elektryczną, a następnie pokazawszy w bieżącym rozdziale istotność tej informacji w procesie podejmowania poprawnych decyzji na rynku energii, wykazujemy tym samym prawdziwość tezy sformułowanej w naszej pracy.

Dlatego w bieżącym rozdziale ilustrujemy wykorzystanie informacji o niepewności prognozy w typowych zadaniach decyzyjnych przy skokowej (dyskretnej) oraz ciągłej zależności skutków decyzji od tejże prognozy. Pokazujemy przy tym, że nieuwzględnienie faktu, iż prognoza jest tylko zmienną losową, a nie informacją pewną, może prowadzić do błędnego wyboru sposobu postępowania w nawet stosunkowo prostych sytuacjach. Jeżeli niepewność przewidywań może prowadzić do zmiany wyboru optymalnego rozwiązania, decyzja ma charakter niestabilny i wymaga uwzględnienia czynnika ryzyka. W tym celu niezbędne jest stworzenie modelu niepewności prognozy oraz oszacowanie jej zachowania.

W ostatnim podrozdziale zajmujemy się bardziej szczegółowo kwestiami wielkości i alokacji zamówienia zakupu energii elektrycznej w warunkach niepewności jej popytu oraz nierównowagi cen zakupu i sprzedaży energii bilansującej, pozwalającej na zrównoważenie zamówienia z zapotrzebowaniem odbiorców. Jak pokazywaliśmy w rozdziale pierwszym, organizacja rynku energii elektrycznej powoduje, że tego rodzaju decyzje muszą być podejmowane z góry, z określonym wyprzedzeniem czasowym. Ostateczne zbilansowanie i kontrakty precyzyjnie dostosowujące posiadane zasoby energii do popytu odbiorców uzyskiwane są w krótkim horyzoncie czasowym rynku dnia następnego lub nawet bieżącego. Stąd znaczenie prognoz krótkoterminowego zapotrzebowania na energię dla tego rodzaju decyzji.

Energia elektryczna jest jednak towarem o ekstremalnie krótkiej trwałości. Brak możliwości jej magazynowania w związku z czynnikiem niepewności zapotrzebowania generuje poważne ryzyko popytowe dla podejmowanych z wyprzedzeniem czasowym decyzji zakupowych, zwłaszcza w sytuacji wystę-

powania nierównowagi cen bilansowania. W rozdziale bieżącym pokazujemy sposób wykorzystania modelu niepewności prognozy do oszacowania tego ryzyka i podjęcia optymalnych decyzji odnośnie do wielkości zakupu i alokacji energii. Wskazujemy przy tym, że brak uwzględnienia tego czynnika dla takiego towaru jak energia elektryczna w analizowanych warunkach prowadzić może do podjęcia niewłaściwych decyzji w tym zakresie.

Zakończenie

Specyfika energii elektrycznej, jej znaczenie dla funkcjonowania współczesnego społeczeństwa, a także brak możliwości magazynowania na skalę przemysłową powodują, że zapotrzebowanie odbiorców tego towaru musi być w każdej chwili dokładnie równoważone przez jego produkcję. Procesy technologiczne w podstawowych źródłach wytwórczych charakteryzują się jednak znaczną bezwładnością czasową, co sprawia, że wolumeny obrotu energią na rynku trzeba ustalać z góry. Transakcje w charakteryzowanych przez nas podstawowych segmentach hurtowego rynku energii elektrycznej, czy to kontraktowym, czy giełdowym, wymagają określenia warunków umów z pewnym wyprzedzeniem czasowym. Nawet jeżeli umowy o dłuższych horyzontach czasowych mogą mieć charakter ramowy, mniej precyzyjny, to w zakresie krótkoterminowym (dnia następnego, z możliwością pewnego skrócenia tego horyzontu do kilku godzin) ich wolumeny określane są precyzyjnie w formie grafików zgłaszanych do realizacji operatorowi systemu przesyłowego, który odpowiada za realizację fizycznych dostaw energii na hurtowym rynku bilansującym.

Decyzje podmiotów działających na rynku energii, których skutki zależą od rzeczywistego wolumenu energii w zawieranych transakcjach, muszą więc być podejmowane z co najmniej dobowym wyprzedzeniem czasowym. Stąd właśnie kluczowe znaczenie dla trafności tychże decyzji krótkoterminowych prognoz zapotrzebowania na energię. Powoduje to konieczność wykorzystania w tej dziedzinie zaawansowanych metod prognostycznych, takich jak sieci neuronowe czy neuronowo-rozmyte, które analizowaliśmy w drugim rozdziale pracy.

Należy jednak pamiętać, że żadna prognoza nie umożliwi dokładnego przewidzenia rzeczywistej wartości prognozowanego zjawiska. Nasza wiedza o przyszłości, nawet w krótkim horyzoncie czasowym, zawsze obarczona jest niepewnością. Na tym polega kolejny specyficzny aspekt handlu energią elektryczną, który stanowi istotny element wspierający tezę naszej pracy. Umowy zawierane na rynku energii niemal nigdy nie są realizowane dokładnie. Niepewność zapotrzebowania przekłada się na niepewność rzeczywistego wolumenu realizowanych transakcji, a co za tym idzie – skutków decyzji, które od niego zależą. Dlatego, zgodnie z tezą tej rozprawy, korzystanie wyłącznie z prognozy określającej wartość oczekiwaną zapotrzebowania na energię może prowadzić do błędnych oszacowań ryzyka popytowego i w konsekwencji do podjęcia nietrafnej decyzji.

Niestety w przypadku złożonych nieliniowych modeli prognostycznych, takich jak sieci neuronowe i neuronowo-rozmyte, nie dysponujemy prostymi metodami oceny niepewności otrzymywanych prognoz. Obecnie stosowane algorytmy wyznaczania wariancji (odchylenia standardowego) wyjścia modelu mają charakter przybliżony albo oparte są na oszacowaniach empirycznych. By wykazać prawdziwość tezy postawionej w tej pracy, niezbędne było więc udowodnienie hipotezy badawczej, w której stwierdziliśmy, że w ogóle istnieją metody z tej dziedziny dające właściwe efekty dla modeli neuronowych i neuronowo-rozmytych w zadaniach prognozy krótkoterminowego zapotrzebowania na energię elektryczną.

Hipoteza ta udowodniona została w rozdziale trzecim rozprawy. Przebadaliśmy w nim szereg obecnie stosowanych metod szacowania wariancji (odchylenia standardowego) wyjścia modeli neuronowych i neuronowo-rozmytych, wykorzystując je w wielu problemach z zakresu prognozowania popytu na energię. Na podstawie wyników tych badań można powiedzieć, że przynajmniej dwie z nich dały właściwe wyniki, pozwalając na poprawną rekonstrukcję rozkładu prawdopodobieństwa zapotrzebowania na energię dla danego wzorca wejściowego prognozy: metoda delta z dokładnym oszacowaniem hesjanu błędu oraz metoda empiryczna oparta na bootstrapie.

Większe znaczenie praktyczne ma, jak się wydaje, pierwsza metoda – oparta na analizie struktury modelu i popełnianego przez niego błędu. Wymaga ona bowiem znacznie niższych nakładów obliczeniowych. Należy jednak podkreślić, że nakłady potrzebne do oszacowania wariancji wyjściowej modelu (odchylenia standardowego) za pomocą bootstrapu nie są na tyle duże, aby przekreślały możliwości praktycznego zastosowania drugiej metody. Oszacowanie empiryczne może być natomiast uważane za bardziej wiarygodne, pewniejsze niż oszacowanie analityczne wykonane przy pewnych założeniach, które spełnione będą tylko w przybliżeniu.

Ponadto pamiętać należy, że obydwie wskazywane metody muszą być weryfikowane dla każdego konkretnego modelu prognostycznego. Wyniki badań prowadzonych na potrzeby naszej pracy wskazują jedynie, że warto rozważyć zastosowanie którejś z nich oraz że te właśnie metody powinny stanowić pierwszy wybór w tej dziedzinie.

Po wykazaniu prawdziwości postawionej hipotezy badawczej i zaprezentowaniu wyników badań świadczących o możliwości zastosowania analizowanych metod w ocenie niepewności prognoz krótkoterminowego zapotrzebowania na energię elektryczną, uzyskanych za pomocą modeli neuronowych i neuronowo-rozmytych, aby ostatecznie dowieść tezy rozprawy, w rozdziale czwartym zaprezentowaliśmy wykorzystanie dodatkowych informacji o rozkładzie zapotrzebowania na energię w najważniejszych typach problemów decyzyjnych. Przedstawiliśmy sposób przełożenia tych informacji na oszacowanie ryzyka decyzji, dowodząc, że posługiwanie się wyłącznie samą wartością prognozy

może w wielu przypadkach prowadzić do wyboru błędnych alternatyw decyzyjnych.

Podsumowując, w naszej pracy pokazaliśmy uwarunkowania związane z krótkoterminową niepewnością popytową na współczesnym rynku energii oraz z jej wpływem na decyzje biznesowe podmiotów działających na rynku. Przebadaliśmy metody szacowania niepewności dla modeli predykcyjnych opartych na sieciach neuronowych i neuronowo-rozmytych, wykazując przydatność tych metod w praktycznych zadaniach z tej dziedziny. Przeanalizowaliśmy następnie sposób zastosowania otrzymanej informacji prognostycznej w podstawowych typach problemów decyzyjnych, koncentrując się przede wszystkim na poprawie jakości procesu decyzyjnego dzięki zastosowanym rozwiązaniom. Przywołane fakty pozwalają nam na stwierdzenie o udowodnieniu tezy postawionej w pracy.

ZAŁĄCZNIK 1

Ważniejsze gradienty i hesjany związane z warstwową siecią perceptronową MLP

Z1.1. Wyznaczanie gradientu błędu sieci MLP względem wag dla danego wzorca treningowego

Gradient błędu sieci przede wszystkim odgrywa kluczową rolę w algorytmie uczenia sieci metodą wstecznej propagacji błędu, więc w praktyce potrzebna nam będzie jego wartość dla danego wzorca treningowego. W naszych rozważaniach ograniczymy się do sieci perceptronowych z jedną warstwą ukrytą i jednym neuronem wyjściowym – postaci dokładnie sprecyzowanej w punkcie 2.2.2. Przytoczmy tu jedynie jeszcze raz podstawowe równanie przetwarzające sieci, które będzie podstawą dalszych obliczeń:

$$y(x_1, x_2, ..., x_n) = \varphi\left(\sum_{i=1}^{h} w_{1i}^{(2)} \varphi\left(\sum_{j=1}^{n} w_{ij}^{(1)} x_j \right) \right) \tag{Z1.1}$$

gdzie h jest liczbą neuronów w warstwie ukrytej, $w_{ij}^{(1)}$ współczynnikiem wagowym j-tego wejścia, i-tego neuronu w warstwie ukrytej, $w_{1i}^{(2)}$ jest współczynnikiem wagowym i-tego wejścia jedynego neuronu wyjściowego, zaś φ funkcją aktywacji neuronów. Oznaczmy jeszcze przez $y_i^{(m)}$, gdzie m = 0, 1, 2, stan (wyjście) i-tego neuronu w warstwie m, przy czym przez m = 0 rozumiemy warstwę wejściową sieci, tzn. $y_i^{(0)} = x_i$, natomiast przez $net_i^{(m)}$, m = 1, 2 – łączne pobudzenie każdego obliczanego neuronu. Wówczas równanie sieci możemy zapisać:

$$\begin{aligned}
y_j^{(0)} &= x_j \qquad , j = 1, ..., n \\
y_i^{(1)} &= \varphi\left(net_i^{(1)}\right) = \varphi\left(\sum_{i=1}^{n} w_{ij}^{(1)} y_j^{(0)} \right) \qquad , i = 1, ..., h \\
y_1^{(2)} &= \varphi\left(net_1^{(2)}\right) = \varphi\left(\sum_{i=1}^{h} w_1^{(2)} y_i^{(1)} \right)
\end{aligned} \tag{Z1.2}$$

Wyznaczymy gradient względem wag, dla błędu kwadratowego sieci:

$$E = \frac{1}{2}\sum_{k=1}^{N}(y_k - y(x_{k1},...,x_{kn}))^2 = \frac{1}{2}\sum_{k=1}^{N}(y_k - y(\mathbf{x}_k))^2 = \sum_{k=1}^{N}\frac{1}{2}e_k^2 = \sum_{k=1}^{N}E_k \tag{Z1.3}$$

gdzie zbiór$\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$ stanowi zbiór danych treningowych, dla których wyznaczono błąd, zaś E_k – część błędu przypadającą na dane odchylenie.

Wyznaczmy pochodną błędu E względem danej wagi. Będzie ona oczywiście równa sumie pochodnych składników błędu E_k, odpowiadających odchyleniom treningowym e_k. Jeżeli operację różniczkowania wykonujemy dodatkowo dla danego wejściowego wzorca treningowego $\mathbf{x}_k$, to zauważymy, że występujące w (3.6.3) wyjścia sieci są stałe, z wyjątkiem jedynego wyjścia zależącego od $\mathbf{x}_k$, czyli $y(\mathbf{x}_k)$. Ich pochodna jest więc równa zero. W konsekwencji, dla danego $\mathbf{x}_k$, gradient błędu E redukuje się tylko do gradientu składnika błędu E_k odpowiadającego tej obserwacji treningowej:

$$\frac{\partial E}{\partial w_{ij}^{(m)}} = \frac{\partial E_k}{\partial w_{ij}^{(m)}} \tag{Z1.4}$$

W dalszej części naszego wyprowadzenia indeks k dla wszystkich pobudzeń i stanów neuronów będziemy dla uproszczenia pomijać, pamiętając jednakże, że ich wartości są wyznaczone dla konkretnego wejściowego wzorca treningowego $\mathbf{x}_k$.

Zauważmy dalej, że odwzorowanie realizowane przez każdy neuron jest złożeniem pobudzenia neuronu oraz funkcji aktywacji. W związku z tym, korzystając ze wzoru na pochodną funkcji złożonej, mamy:

$$\frac{\partial E_k}{\partial w_{ij}^{(m)}} = \frac{\partial E_k}{\partial net_i^{(m)}}\frac{\partial net_i^{(m)}}{\partial w_{ij}^{(m)}} \tag{Z1.5}$$

Pobudzenie neuronu jest funkcją liniową, więc

$$\frac{\partial E_k}{\partial w_{ij}^{(m)}} = \frac{\partial E_k}{\partial net_i^{(m)}}\frac{\partial}{\partial w_{ij}^{(m)}}\left(\sum_k w_{ik}^{(m)} y_k^{(m-1)}\right) = \frac{\partial E_k}{\partial net_i^{(m)}} y_j^{(m-1)} \tag{Z1.6}$$

Oznaczmy teraz przez $\delta_i^{(m)}$ pochodną błędu sieci względem pobudzenia i-tego neuronu sieci w warstwie m:

$$\delta_i^{(m)} = -\frac{\partial E_k}{\partial net_i^{(m)}} \tag{Z1.7}$$

Wówczas (Z1.6) możemy zapisać:

$$\frac{\partial E_k}{\partial w_{ij}^{(m)}} = -\delta_i^{(m)} y_j^{(m-1)} \tag{Z1.8}$$

czyli pochodna błędu względem każdej z wag może zostać wyznaczona jako iloczyn pochodnej błędu względem pobudzenia przez wartość wejścia związanego z tą wagą. Jej znalezienie wymaga wobec tego określenia wartości $\delta_i^{(m)}$.

Zauważmy dalej, że pochodną błędu względem pobudzenia każdego neuronu można rozłożyć na pochodną błędu względem jego wyjścia i pochodną wyjścia neuronu względem pobudzenia. Mamy więc:

$$\delta_i^{(m)} = -\frac{\partial E_k}{\partial net_i^{(m)}} = -\frac{\partial E_k}{\partial y_i^{(m)}} \frac{\partial y_i^{(m)}}{\partial net_i^{(m)}} = -\frac{\partial E_k}{\partial y_i^{(m)}} \varphi'(net_i^{(m)}) \tag{Z1.9}$$

co natychmiast daje nam wartość błędu δ dla neuronu wyjściowego:

$$\begin{aligned} \delta_1^{(2)} &= -\frac{\partial E_k}{\partial y_1^{(2)}} \varphi'(net_1^{(2)}) = -\frac{\partial}{\partial y(\mathbf{x}_k)} \left(\frac{1}{2} (y_k - y(\mathbf{x}_k))^2 \right) \varphi'(net_1^{(2)}) = \\ &= (y_k - y(\mathbf{x}_k)) \varphi'(net_1^{(2)}) \end{aligned} \tag{Z1.10}$$

Dla neuronów warstwy ukrytej zastosujemy podobną procedurę jak w przypadku (Z1.6). Zauważmy, że stan (wyjście) każdego neuronu warstwy ukrytej stanowi wejście neuronu wyjściowego. Pochodną błędu względem stanu neuronu warstwy ukrytej możemy więc rozłożyć na pochodną błędu względem pobudzenia neuronu wyjściowego i pochodną pobudzenia neuronu wyjściowego względem wejścia neuronu:

$$\begin{aligned} \frac{\partial E_k}{\partial y_i^{(1)}} &= \frac{\partial E_k}{\partial net_1^{(2)}} \frac{\partial net_1^{(2)}}{\partial y_i^{(1)}} = \frac{\partial E_k}{\partial net_1^{(2)}} \frac{\partial}{\partial y_i^{(1)}} \left(\sum_k w_{1k}^{(2)} y_k^{(1)} \right) = \\ &= \frac{\partial E_k}{\partial net_1^{(2)}} w_{1i}^{(2)} = -\delta_1^{(2)} w_{1i}^{(2)} \end{aligned} \tag{Z1.11}$$

Podstawiając (Z1.11) do (Z1.9), otrzymujemy:

$$\delta_i^{(1)} = -\frac{\partial E_k}{\partial y_i^{(m)}} \varphi'(net_i^{(m)}) = \delta_1^{(2)} w_{1i}^{(2)} \varphi'(net_i^{(1)}) \tag{Z1.12}$$

Mamy już wszystko, co potrzebne do wyznaczenia gradientu (pochodnych) błędu naszej sieci MLP (Z1.1) względem wag. Zauważmy, że udało nam się uzależnić błędy $\delta_i^{(1)}$ dla neuronów warstwy ukrytej od błędów $\delta_1^{(2)}$ dla neuronów (czy też właściwie neuronu) warstwy wyjściowej. Wymusza to pewien uporządkowany sposób wyznaczania pochodnych polegający na wykonaniu obliczeń wstecz – zaczynamy od neuronu wyjściowego i posuwamy się w stronę neuronów wejściowych (stąd właśnie wzięła się nazwa metody uczenia sieci neuronowej: algorytm wstecznej propagacji błędu). Dokładniej proces wyznaczania gradientu błędu dla danej obserwacji treningowej $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$ przyjmuje postać:

1. Podajemy wzorzec $\mathbf{x}_k$ na wejście sieci. Przy wykorzystaniu zależności (3.6.1) (lub (Z1.2)) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Znajdujemy przy użyciu (Z1.10) współczynnik $\delta_1^{(2)}$ dla neuronu warstwy wyjściowej, a następnie za pomocą (Z1.12) oraz obliczonego $\delta_1^{(2)}$ – błędy $\delta_i^{(1)}$ dla wszystkich neuronów warstwy ukrytej.

3. Korzystając z (Z1.8) i obliczonych błędów δ, obliczamy pochodne względem wszystkich wag sieci, pamiętając oczywiście, że $y_i^{(0)} = x_{ki}$.

W przypadku sieci o większej liczbie warstw ukrytych sposób ostatecznego wyznaczania pochodnych w punkcie 3 i wzór (Z1.8) pozostają bez zmian. Różnica będzie polegała na wyznaczeniu w kroku wstecznym dwóch błędów $\delta_i^{(m)}$, dla neuronów kolejnych warstw (licząc od strony wyjścia sieci). Dla pierwszych dwóch warstw, tzn. wyjściowej i ostatniej warstwy ukrytej, obliczenia przebiegają tak samo jak w przypadku sieci (3.6.1). Idąc dalej w stronę wejść, dla kolejnych warstw ukrytych będziemy dalej stosować łańcuchową metodę obliczania $\delta_i^{(m)}$ za pomocą błędów w warstwie następnej $\delta_i^{(m+1)}$, podobną do (Z1.12); musimy tylko uwzględnić fakt, że w warstwie m+1 jest kilka neuronów, a nie jeden. Jeżeli oznaczymy przez h_m liczbę neuronów w każdej warstwie sieci, to w związku z tym mamy:

$$\frac{\partial E_k}{\partial y_i^{(m)}} = \sum_{t=1}^{h_{m+1}} \frac{\partial E_k}{\partial net_t^{(m+1)}} \frac{\partial net_t^{(m+1)}}{\partial y_i^{(m)}} =$$
$$= \sum_{t=1}^{h_{m+1}} \frac{\partial E_k}{\partial net_t^{(m+1)}} \frac{\partial}{\partial y_i^{(m)}} \left(\sum_j w_{tj}^{(m+1)} y_j^{(m)} \right) = \tag{Z1.13}$$
$$= \sum_{t=1}^{h_{m+1}} \frac{\partial E_k}{\partial net_t^{(m+1)}} w_{ti}^{(m+1)} = -\sum_{t=1}^{h_{m+1}} \delta_t^{(m+1)} w_{ti}^{(m+1)}$$

a co za tym idzie:

$$\delta_i^{(m)} = \varphi'(net_i^{(m)}) \sum_{t=1}^{h_{m+1}} \delta_t^{(m+1)} w_{ti}^{(m+1)} \tag{Z1.14}$$

Z1.2. Wyznaczanie hesjanu błędu sieci MLP względem wag

Zastanówmy się obecnie nad wyznaczeniem hesjanu błędu sieci neuronowej względem wag. Wyznaczenie drugich pochodnych błędu, które stanowią elementy macierzy hesjanu, ma istotne znaczenie zarówno dla niektórych bardziej zaawansowanych algorytmów uczenia (bądź douczania) sieci, jak i dla optymalizacji jej struktury (Bishop 1995). Zagadnienia te wykraczają jednak tematycznie poza zakres niniejszej pracy i nie są w niej omawiane. W naszym przypadku odwrotność hesjanu błędu wykorzystywana jest w metodzie delta do oszacowania macierzy kowariancji wag sieci i wyznaczenia odchylenia standardowego (wariancji) warunkowego rozkładu wartości wyjścia modelu (otrzymywanej prognozy) dla danego wzorca wejściowego (patrz punkt 3.3.2).

W bieżącym załączniku zajmujemy się wyłącznie dokładnym wyznaczaniem hesjanu błędu dla konkretnej architektury sieci neuronowej – warstwowych perceptronów MLP. Istnieje szereg metod przybliżonego szacowania hesjanu, które mają charakter uniwersalny i mogą być stosowane (przynajmniej do pewnego miejsca) dla modeli różnego typu. W interesującym nas kontekście wyznaczania macierzy kowariancji wag modelu najważniejszą z nich jest metoda aproksymacji iloczynem skalarnym (albo aproksymacja Levenberga–Marquarda), którą przedstawiamy w punkcie 3.3.2.

Oznaczmy macierz hesjanu przez $\mathbf{H} = [h_{qr}]$, q, $r = 1, \ldots, p$, gdzie p jest równe liczbie współczynników wagowych sieci. Pozostałe oznaczenia stosowane w tym punkcie są takie same jak w punkcie poprzednim, Z1.1, chyba że zostanie to jasno powiedziane inaczej. Nasz cel polega na znalezieniu drugich pochodnych dla dowolnej pary wag:

$$\frac{\partial^2 E}{\partial w_{ij}^{(m)} \partial w_{lt}^{(s)}} \tag{Z1.15}$$

gdzie, podobnie jak w poprzednim punkcie, $w_{ij}^{(m)}$ jest wagą j-tego wejścia, i-tego neuronu w m-tej warstwie. Zauważmy jeszcze, że hesjan **H** jest macierzą symetryczną, ponieważ oczywiście mamy:

$$\frac{\partial^2 E}{\partial w_{ij}^{(m)} \partial w_{lt}^{(s)}} = \frac{\partial^2 E}{\partial w_{lt}^{(s)} \partial w_{ij}^{(m)}} \tag{Z1.16}$$

Jeżeli więc założymy sobie jakieś uporządkowanie wag występujących w wyznaczanej drugiej pochodnej względem siebie, np. związane z położeniem w strukturze sieci, to nie zmniejszamy ogólności rozważań. Pochodna dla odwrotnego układu wag jest równa wyznaczonej.

Szczegółowy algorytm przedstawimy dla sieci z jednym neuronem wyjściowym i jedną warstwą ukrytą i h neuronach (równanie (Z1.1) lub (Z1.2)), ale rozpoczniemy od jego wyprowadzenia w przypadku ogólniejszym, dla dowolnej liczby M warstw ukrytych, z których każda ma h_m neuronów. Jak zobaczymy, metoda ta opiera się na zbliżonych procedurach rekursywnego przechodzenia przez kolejne warstwy sieci jak w przypadku algorytmu wstecznej propagacji błędu (wyznaczania gradientu błędu względem wag). Rozważania w bieżącym punkcie opierają się na algorytmie wyznaczania hesjanu sieci MLP przedstawionym przez Bishopa (Bishop 1992; 1995). Podobne rozwiązania rozważali Buntine i Weigand (1994).

Przyjmijmy, tak samo jak w poprzednim punkcie, że sieć została nauczona na zbiorze danych $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$, więc różniczkowany błąd kwadratowy E wyznaczono dla tych właśnie wzorców. Również w tym przypadku możemy zgodnie ze wzorem (Z1.3) przedstawić błąd E jako sumę błędów E_k przypadających na odchylenia poszczególnych obserwacji treningowych. Obecnie jednak interesuje nas wyznaczenie pochodnych funkcji błędu dla całego zbioru treningowego, a nie dla pojedynczego wzorca. Każdy element hesjanu będzie więc równy pochodnym składników błędu E_k, zsumowanym po wszystkich wzorcach treningowych.

$$\frac{\partial^2 E}{\partial w_{ij}^{(m)} \partial w_{lt}^{(s)}} = \sum_{k=1}^{N} \frac{\partial^2 E_k}{\partial w_{ij}^{(m)} \partial w_{lt}^{(s)}} \tag{Z1.17}$$

Zajmijmy się więc oszacowaniem drugiej pochodnej składnika pojedynczego błędu E_k. Podobnie jak w poprzednim punkcie, indeks k dla wszystkich

pobudzeń i stanów neuronów będziemy dla uproszczenia pomijać, pamiętając jednakże, że ich wartości są wyznaczone dla konkretnego wejściowego wzorca treningowego $\mathbf{x}_k$.

W tym miejscu właśnie poczynimy pewne założenie. Załóżmy mianowicie, że jeżeli, jak w (Z1.17), wyznaczymy pierwszą pochodną względem wagi $w_{lt}^{(s)}$, a następnie różniczkujemy tę pochodną względem $w_{ij}^{(m)}$, to $m \leq s$, czyli drugi w kolejności różniczkowania neuron i jest w tej samej lub wcześniejszej (położonej bliżej wejść) warstwie sieci neuronowej, co pierwszy neuron l. Takie założenie o uporządkowaniu wag nie zmniejsza oczywiście ogólności naszych rozważań, ponieważ macierz hesjanu jest symetryczna. W przypadku pochodnych, dla których nie jest ono spełnione, czyli $s > m$, odwracamy kolejność różniczkowania, a więc najpierw obliczamy pochodną względem $w_{ij}^{(m)}$, następnie względem $w_{lt}^{(s)}$. Wynik będzie taki sam.

Tak naprawdę wystarczyłoby nam nieco słabsze założenie, tzn. aby druga w kolejności różniczkowania waga $w_{ij}^{(m)}$ nie leżała w grafie tworzonym przez węzły sieci, na ścieżkach łączących neuron l (do którego należy waga $w_{lt}^{(s)}$) z wyjściem sieci, co, naturalnie, jest spełnione, jeśli $m \leq s$. Wówczas bowiem możemy napisać:

$$\frac{\partial^2 E_k}{\partial w_{ij}^{(m)} \partial w_{lt}^{(s)}} = \frac{\partial}{\partial net_i^{(m)}} \left(\frac{\partial E_k}{\partial w_{lt}^{(s)}} \right) \frac{\partial net_i^{(m)}}{\partial w_{ij}^{(m)}} = \frac{\partial}{\partial net_i^{(m)}} \left(\frac{\partial E_k}{\partial w_{lt}^{(s)}} \right) y_j^{(m-1)} \tag{Z1.18}$$

Zauważmy, że pierwsza równość w powyższym wzorze zakłada, iż pochodna błędu E_k względem $w_{lt}^{(s)}$ zależna jest od wagi $w_{ij}^{(m)}$ wyłącznie za pośrednictwem pobudzenia $net_i^{(m)}$. Przy tym na podstawie (Z1.8) pochodna ta jest równa:

$$\frac{\partial E_k}{\partial w_{lt}^{(s)}} = -\delta_l^{(s)} y_t^{(s-1)} \tag{Z1.19}$$

Wielkości $\delta_l^{(s)}$ znamy z algorytmu wstecznej propagacji błędów z poprzedniego punktu. Przypomnijmy, że obliczane są one ze wzoru (Z1.14), tj. za pomocą łańcuchowej procedury rekursywnej – wykorzystuje się w niej m.in. wszystkie wagi $w_{tl}^{(s+1)}$, przez które przepływa sygnał wyjściowy neuronu l, oraz błędy $\delta_t^{(s+1)}$ w kolejnej warstwie. Ciągnąc dalej ten łańcuch wnioskowania, wielkość $\delta_l^{(s)}$, a co za tym idzie, pierwsza pochodna błędu względem $w_{lt}^{(s)}$ zależy bezpośrednio od wszystkich wag na ścieżkach między neuronem l a wyjściem sieci. Gdyby więc nasze założenie nie było spełnione, czyli waga $w_{ij}^{(m)}$ leżała na tej ścieżce, to pierwsze przekształcenie w (Z1.18) byłoby nieprawidłowe.

Idąc dalej, wykorzystajmy zależność (Z1.19) w (Z1.18):

$$\begin{aligned}\frac{\partial^2 E_k}{\partial w_{ij}^{(m)} \partial w_{lt}^{(s)}} &= \frac{\partial}{\partial net_i^{(m)}} \left(-\delta_l^{(s)} y_t^{(s-1)} \right) y_j^{(m-1)} = \\ &= -y_j^{(m-1)} \left(\delta_l^{(s)} \frac{\partial y_t^{(s-1)}}{\partial net_i^{(m)}} + y_t^{(s-1)} \frac{\partial \delta_l^{(s)}}{\partial net_i^{(m)}} \right) = \\ &= -y_j^{(m-1)} \left(\delta_l^{(s)} \frac{\partial \left(\varphi(net_t^{(s-1)}) \right)}{\partial net_i^{(m)}} + y_t^{(s-1)} \frac{\partial \delta_l^{(s)}}{\partial net_i^{(m)}} \right) = \\ &= -y_j^{(m-1)} \delta_l^{(s)} \varphi'(net_t^{(s-1)}) \frac{\partial net_t^{(s-1)}}{\partial net_i^{(m)}} - y_j^{(m-1)} y_t^{(s-1)} \frac{\partial \delta_l^{(s)}}{\partial net_i^{(m)}}\end{aligned} \tag{Z1.20}$$

Wprowadźmy dwie dodatkowe zmienne:

$$a_{ti}^{(s-1)m} = \frac{\partial net_t^{(s-1)}}{\partial net_i^{(m)}} \tag{Z1.21}$$

$$b_{li}^{sm} = \frac{\partial \delta_l^{(s)}}{\partial net_i^{(m)}} \tag{Z1.22}$$

Stosując te oznaczenia, możemy (Z1.20) zapisać w następujący sposób:

$$\frac{\partial^2 E_k}{\partial w_{ij}^{(m)} \partial w_{lt}^{(s)}} = -y_j^{(m-1)} \delta_l^{(s)} \varphi'(net_t^{(s-1)}) a_{ti}^{(s-1)m} - y_j^{(m-1)} y_t^{(s-1)} b_{li}^{sm} \tag{Z1.23}$$

Nasza formuła dla drugich pochodnych błędu sieci wymaga jeszcze wyznaczenia występujących w (Z1.20) wielkości $a_{ti}^{(s-1)m}$ i b_{li}^{sm}. Rozpocznijmy od wielkości pochodnej pobudzenia neuronu t względem pobudzenia neuronu i, oznaczonej przez $a_{ti}^{(s-1)m}$. Jeżeli spojrzymy na zależność (Z1.21), to widzimy od razu, że:

– jeżeli obydwa neurony się pokrywają, tzn. $t = i$ oraz $s - 1 = m$, to pochodna pobudzenia względem siebie jest równa jeden, czyli $a_{tt}^{mm} = 1$,

– jeżeli neuron t nie leży na ścieżce propagacji sygnału przez sieć, wychodzącej z neuronu i, to znaczy w naszym przypadku, że $s - 1 < m$ lub $s - 1 = m$ i $t \neq i$, czyli pobudzenie neuronu t nie zależy od pobudzenia neuronu i, a więc ma charakter stały, pochodna jest równa zero, $a_{ti}^{(s-1)m} = 0$,

– jeżeli nie zachodzą oba poprzednie przypadki, to znaczy neuron t leży w którejś z kolejnych warstw przed neuronem i ($s - 1 > m$), musimy wartość $a_{ti}^{(s-1)m}$ obliczyć, korzystając z następującej rekursywnej reguły:

$$a_{zi}^{cm} = \frac{\partial net_z^{(c)}}{\partial net_i^{(m)}} = \sum_{r=1}^{h_{c-1}} \frac{\partial net_z^{(c)}}{\partial net_r^{(c-1)}} \frac{\partial net_r^{(c-1)}}{\partial net_i^{(m)}} = \\ = \sum_{r=1}^{h_{c-1}} w_{zr}^{(c)} \varphi'(net_r^{(c-1)}) a_{ri}^{(c-1)m} \tag{Z1.24}$$

Z reguły (Z1.24) korzystamy na zasadzie propagacji do przodu: dla warstwy m ustawiamy wartości a_{ri}^{mm} na zero lub na jeden zgodnie z zasadami z pierwszych dwóch punktów, a następnie korzystamy z (Z1.24) dla kolejnych warstw aż do obliczenia $a_{ti}^{(s-1)m}$.

Dosyć podobnie, tylko na zasadach propagacji wstecz, poradzimy sobie z obliczeniami pochodnych błędów δ względem pobudzeń neuronów, tj. b_{li}^{sm}. Przypomnijmy, że w punkcie Z1.1 znaleźliśmy rekurencyjną formułę wyznaczania δ, w zależności od błędów neuronów w następnej warstwie (wzór (Z1.14)). W naszym przypadku będziemy więc mieli:

$$\delta_l^{(s)} = \varphi'(net_l^{(s)}) \sum_{r=1}^{h_{s+1}} \delta_r^{(s+1)} w_{rl}^{(s+1)} \tag{Z1.25}$$

Podstawiając zatem (Z1.25) do definicji b_{li}^{sm} (Z1.22), otrzymujemy:

$$b_{li}^{sm} = \frac{\partial}{\partial net_i^{(m)}} \left(\varphi'(net_l^{(s)}) \sum_{r=1}^{h_{s+1}} \delta_r^{(s+1)} w_{rl}^{(s+1)} \right) \tag{Z1.26}$$

Korzystając teraz ze wzoru na pochodną iloczynu, a następnie pochodną funkcji złożonej, otrzymujemy:

$$b_{li}^{sm} = \frac{\partial}{\partial net_i^{(m)}} \left(\varphi'(net_l^{(s)}) \right) \sum_{r=1}^{h_{s+1}} \delta_r^{(s+1)} w_{rl}^{(s+1)} + \\ + \varphi'(net_l^{(s)}) \frac{\partial}{\partial net_i^{(m)}} \left(\sum_{r=1}^{h_{s+1}} \delta_r^{(s+1)} w_{rl}^{(s+1)} \right) = \\ = \varphi''(net_l^{(s)}) \frac{\partial net_l^{(s)}}{\partial net_i^{(m)}} \sum_{r=1}^{h_{s+1}} \delta_r^{(s+1)} w_{rl}^{(s+1)} + \varphi'(net_l^{(s)}) \sum_{r=1}^{h_{s+1}} w_{rl}^{(s+1)} \frac{\partial \delta_r^{(s+1)}}{\partial net_i^{(m)}} \tag{Z1.27}$$

Zastępując w (Z1.27) wartość a_{li}^{sm} (na podstawie definicji (Z1.21)) oraz $b_{ri}^{(s+1)m}$ (definicja (Z1.22)), otrzymujemy ostateczną formułę rekurencyjną na wsteczną propagację pochodnych błędów δ względem pobudzeń neuronów:

$$b_{li}^{sm} = \varphi''(net_l^{(s)}) a_{li}^{sm} \sum_{r=1}^{h_{s+1}} \delta_r^{(s+1)} w_{rl}^{(s+1)} + \varphi'(net_l^{(s)}) \sum_{r=1}^{h_{s+1}} w_{rl}^{(s+1)} b_{ri}^{(s+1)m} \qquad \text{(Z1.28)}$$

Obliczenia w (Z1.28) przebiegają po wszystkich neuronach w warstwie następnej po s, należy więc rozpocząć je od wyjściowych węzłów sieci i za pomocą powyższej formuły kontynuować je wstecz aż do obliczenia b_{li}^{sm}.

Jeszcze raz przypomnijmy, że pochodna względem pobudzenia $net_i^{(m)}$ w (Z1.26) wynika z pochodnej względem wagi $w_{ij}^{(m)}$ (patrz wzór (Z1.18)). Przedstawione obliczenia są więc poprawne jedynie przy założeniu że waga $w_{ij}^{(m)}$ nie pojawia się bezpośrednio w ciągach obliczeń rekurencyjnych w (Z1.28). Przypomnijmy jednak, że możemy przyjąć założenie, odnośnie do takiej kolejności obliczeń elementów hesjanu, że $m \le s$. Macierz hesjanu jest symetryczna, więc założenie to, jak już mówiliśmy wcześniej, nie zmniejsza ogólności rozważań.

Rekurencja (Z1.28) wymaga jeszcze określenia wartości początkowych, tzn. ustalenia wartości $b_{li}^{(wy)m}$ dla neuronów wyjściowych (lub w naszym przypadku jednego neuronu wyjściowego). Z definicji $b_{li}^{(wy)m}$ oraz ze wzoru (Z1.10) na błędy neuronów wyjściowych możemy napisać:

$$\begin{aligned} b_{li}^{(wy)m} &= \frac{\partial \delta_l^{(wy)}}{\partial net_i^{(m)}} = \frac{\partial}{\partial net_i^{(m)}} \left(-\varphi'(net_l^{(wy)}) \frac{\partial E_k}{\partial y_l^{(wy)}} \right) = \\ &= -\varphi''(net_l^{(wy)}) \frac{\partial net_l^{(wy)}}{\partial net_i^{(m)}} \frac{\partial E_k}{\partial y_l^{(wy)}} - \varphi'(net_l^{(wy)}) \frac{\partial^2 E_k}{\partial y_l^{(wy)} \partial y_l^{(wy)}} \frac{\partial y_l^{(wy)}}{\partial net_i^{(m)}} \\ &= -\varphi''(net_l^{(wy)}) \frac{\partial net_l^{(wy)}}{\partial net_i^{(m)}} \frac{\partial E_k}{\partial y_l^{(wy)}} - \left(\varphi'(net_l^{(wy)})\right)^2 \frac{\partial^2 E_k}{\partial y_l^{(wy)} \partial y_l^{(wy)}} \frac{\partial net_l^{(wy)}}{\partial net_i^{(m)}} \end{aligned} \qquad \text{(Z1.29)}$$

Z definicji $a_{li}^{(wy)m}$ (Z1.21) otrzymujemy:

$$b_{li}^{(wy)m} = -a_{li}^{(wy)m} \left(\varphi''(net_l^{(wy)}) \frac{\partial E_k}{\partial y_l^{(wy)}} + \left(\varphi'(net_l^{(wy)})\right)^2 \frac{\partial^2 E_k}{\partial y_l^{(wy)} \partial y_l^{(wy)}} \right) \qquad \text{(Z1.30)}$$

Zauważmy, że człon (Z1.30) ujęty w nawiasy nie zależy od indeksu drugiego neuronu, względem którego wyznaczamy hesjan, tj. od i. Można go więc obliczyć tylko raz, dla neuronu wyjściowego. Oznaczmy go przez $\mathbf{H}_l$. Ostatecznie wartość początkową $b_{li}^{(wy)m}$ możemy przedstawić jako:

$$b_{li}^{(wy)m} = -a_{li}^{(wy)m} \mathbf{H}_l \qquad \text{(Z1.31a)}$$

gdzie

$$\mathbf{H}_l = \varphi''(net_l^{(wy)})\frac{\partial E_k}{\partial y_l^{(wy)}} + \left(\varphi'(net_l^{(wy)})\right)^2 \frac{\partial^2 E_k}{\partial y_l^{(wy)} \partial y_l^{(wy)}} = \\ = \frac{\partial^2 E_k}{\partial net_l^{(wy)} \partial net_l^{(wy)}} \tag{Z1.31b}$$

Podsumujmy naszą dyskusję odnośnie do obliczeń hesjanu błędu dla dowolnej warstwowej sieci perceptronowej MLP. Schemat postępowania możemy naszkicować następująco (Bishop 1995):

1. Dla danego treningowego wzorca wejściowego $\mathbf{x}_k$ przeliczamy wszystkie neurony sieci, znajdując ich stany (wyjścia) oraz pobudzenia. Równolegle do tych obliczeń możemy dla każdej pary węzłów sieci wyznaczyć niezerowe wartości $a_{ti}^{(s-1)m}$, odpowiednio je inicjując i w razie potrzeby obliczając przez propagację do przodu (Z1.24).

3. Stosując standardowe formuły z metody wstecznej propagacji, wyznaczamy błędy neuronów wyjściowych $\delta_l^{(wy)}$ oraz $\mathbf{H}_l$ i $b_{li}^{(wy)m}$ (korzystając ze wzorów (Z1.31)).

4. Korzystamy ze standardowych zależności propagacji błędu wstecz i wyznaczamy błędy δ wszystkich neuronów sieci; ponadto, stosując (Z1.28), dokonujemy propagacji wstecz b_{li}^{sm}.

5. Za pomocą wzoru (Z1.23) obliczamy cząstkowe elementy macierzy hesjanu błędu **H** dla danego wzorca treningowego. Zwracamy uwagę, by wyznaczać tylko właściwą połówkę **H**, tak by drugie różniczkowanie odbywało się dla wagi z warstwy nieleżącej dalej niż pierwsza, tj. $m \le s$; wagi nad główną przekątną wynikają z symetrycznego charakteru macierzy **H**.

6. Sumujemy obliczone wcześniej elementy cząstkowe macierzy hesjanu **H** po wszystkich wzorcach treningowych.

Przedstawiony algorytm upraszcza się znacznie dla modelu sieci MLP z jedną warstwą ukrytą i jednym neuronem wyjściowym – taki wykorzystywaliśmy w krótkoterminowych prognozach zapotrzebowania na energię w rozdziale 2 (wzór (Z1.1) lub (Z1.2)). Na podstawie podanych zależności elementy cząstkowe macierzy hesjanu błędu dla danego wzorca treningowego wyznaczyć możemy w trzech następujących krokach.

1. Obie wagi są wagami neuronu wyjściowego

Na podstawie ogólnego równania pozwalającego na obliczenie elementów hesjanu błędu, danego przez (Z1.23), dla drugiej pochodnej względem wag neuronu wyjściowego otrzymujemy następującą zależność:

$$\frac{\partial^2 E_k}{\partial w_{1j}^{(2)} \partial w_{1t}^{(2)}} = -y_j^{(1)} \delta_1^{(2)} \varphi'(net_t^{(1)}) a_{t1}^{(1)(2)} - y_j^{(1)} y_t^{(1)} b_{11}^{(2)(2)} \tag{Z1.32}$$

Zauważmy jednak, że zgodnie z zasadami wyznaczania $a_{ti}^{(s-1)m}$, określonymi w drugim punkcie warunków początkowych zależności (Z1.24), $a_{t1}^{(1)(2)} = 0$, ponieważ pobudzenie $net_t^{(1)}$ dotyczy neuronu leżącego we wcześniejszej warstwie niż neuronu wyjściowego związanego z pobudzeniem $net_1^{(2)}$. Ponadto wartość $b_{11}^{(2)(2)}$ dla neuronu warstwy wyjściowej możemy wyznaczyć bezpośrednio z (Z1.31) jako:

$$b_{11}^{(2)(2)} = -a_{11}^{(2)(2)} \mathbf{H}_1 = -\mathbf{H}_1 \tag{Z1.33}$$

jako że $a_{t1}^{(2)(2)} = 1$ na mocy pierwszego punktu warunków początkowych dla (Z1.24), gdyż chodzi o wagi tego samego neuronu. Ostatecznie więc element cząstkowy hesjanu błędu dla wag neuronu wyjściowego możemy zapisać w następującej postaci:

$$\frac{\partial^2 E_k}{\partial w_{1j}^{(2)} \partial w_{1t}^{(2)}} = y_j^{(1)} y_t^{(1)} \mathbf{H}_1 \tag{Z1.34}$$

Wartość $\mathbf{H}_1$ wyznaczamy ze wzoru (Z1.31b).

2. Obie wagi są wagami warstwy ukrytej

Ponownie korzystając z ogólnego równania dla elementów hesjanu błędu, danego przez (Z1.23), dla drugiej pochodnej wag względem tej samej warstwy ukrytej, otrzymamy:

$$\frac{\partial^2 E_k}{\partial w_{ij}^{(1)} \partial w_{lt}^{(1)}} = -y_j^{(0)} \delta_l^{(1)} \varphi'(net_t^{(0)}) a_{ti}^{(0)(1)} - y_j^{(0)} y_t^{(0)} b_{li}^{(1)(1)} \tag{Z1.35}$$

gdzie przez $\varphi'(net_t^{(0)})$ rozumiemy pochodną pobudzenia neuronu wejściowego związanego ze zmienną objaśniającą x_t. Neurony warstwy wejściowej jedynie przekazują dane do dalszych warstw sieci, ich funkcja aktywacji ma więc charakterystykę liniową, a jej pochodna równa się 1. Nie jest to zresztą takie istotne, ponieważ, podobnie jak w poprzednim punkcie, $a_{ti}^{(0)(1)} = 0$ na mocy drugiego punktu warunków początkowych zależności (Z1.24), dla wyznaczania $a_{ti}^{(s-1)m}$.

Stosując dalej rekurencyjną formułę (Z1.28) do wyznaczania $b_{li}^{(1)(1)}$, otrzymujemy następującą zależność:

$$\frac{\partial^2 E_k}{\partial w_{ij}^{(1)} \partial w_{lt}^{(1)}} = -y_j^{(0)} y_t^{(0)} b_{li}^{(1)(1)} = \\ = -y_j^{(0)} y_t^{(0)} \left(\varphi''(net_l^{(1)}) a_{li}^{(1)(1)} \sum_{r=1}^{h_2} \delta_r^{(2)} w_{rl}^{(2)} + \varphi'(net_l^{(1)}) \sum_{r=1}^{h_2} w_{rl}^{(2)} b_{ri}^{(2)(1)} \right) \tag{Z1.36}$$

Zauważmy teraz, że mamy jeden neuron w warstwie wyjściowej, więc $h_2 = 1$. Następnie wartości $a_{li}^{(1)(1)}$ dla neuronów tej samej warstwy są równe 1, jeżeli $l = i$, 0 w przeciwnym przypadku (patrz pierwsze dwa punkty warunków początkowych zależności (Z1.24) dla wyznaczania $a_{ti}^{(s-1)m}$). A zatem $a_{li}^{(1)(1)} = \Delta_{li}$, gdzie Δ_{li} jest deltą Kroneckera. Dalej

$$b_{1i}^{(2)(1)} = -a_{1i}^{(2)(1)} \mathbf{H}_1 = -\mathbf{H}_1 \sum_{r=1}^{h_1} w_{1r}^{(2)} \varphi'(net_r^{(1)}) a_{ri}^{(1)(1)} = \\ = -\mathbf{H}_1 \sum_{r=1}^{h_1} w_{1r}^{(2)} \varphi'(net_r^{(1)}) \Delta_{ri} = -\mathbf{H}_1 w_{1i}^{(2)} \varphi'(net_i^{(1)}) \tag{Z1.37}$$

Wykorzystując powyższe informacje w (Z1.36), otrzymujemy ostatecznie zależność dla drugich pochodnych względem wag warstwy ukrytej:

$$\frac{\partial^2 E_k}{\partial w_{ij}^{(1)} \partial w_{lt}^{(1)}} = \\ = -y_j^{(0)} y_t^{(0)} \left(\varphi''(net_l^{(1)}) \Delta_{li} \delta_1^{(2)} w_{1l}^{(2)} - \varphi'(net_l^{(1)}) \varphi'(net_i^{(1)}) w_{1l}^{(2)} w_{1i}^{(2)} \mathbf{H}_1 \right) \tag{Z1.38}$$

gdzie, przypomnijmy, $\delta_1^{(2)}$ jest standardowym błędem neuronu wyjściowego z metody wstecznej propagacji błędu (wzór (Z1.10)), zaś wartość $\mathbf{H}_1$ wyznaczamy ze wzoru (Z1.31b). Ze względów efektywności obliczeń pamiętać powinniśmy o symetryczności hesjanu, więc, jeśli zastosujemy (Z1.38) do obliczenia pochodnej błędu względem pary wag $w_{ij}^{(1)}$, $w_{lt}^{(1)}$, to pochodna względem odwrotnego układu $w_{lt}^{(1)}$, $w_{ij}^{(1)}$ będzie taka sama.

3. Pierwsza waga w warstwie wyjściowej, druga w warstwie ukrytej

Zauważmy, że kolejność warstw jest istotna, ponieważ nasze wzory wyprowadzane były przy założeniu, że $m \leq s$. Gdyby pierwsza waga była z warstwy ukrytej, a druga – z wyjściowej, musielibyśmy korzystać z symetrii macierzy hesjanu.

Ponownie skorzystajmy z ogólnego wzoru na wyrazy macierzy drugiej pochodnej względem wag (Z1.23):

$$\frac{\partial^2 E_k}{\partial w_{ij}^{(1)} \partial w_{1t}^{(2)}} = -y_j^{(0)} \delta_1^{(2)} \varphi'(net_t^{(1)}) a_{ti}^{(1)(1)} - y_j^{(0)} y_t^{(1)} b_{1i}^{(2)(1)} \tag{Z1.39}$$

Podobnie jak w poprzednim punkcie, skorzystamy z faktu, że na podstawie dwóch pierwszych punktów warunków początkowych zależności (Z1.24) dla wyznaczania $a_{ti}^{(s-1)m}$ mamy $a_{ti}^{(1)(1)} = \Delta_{ti}$, gdzie Δ_{ti} jest deltą Kroneckera. Ponadto wartość $b_{1i}^{(2)(1)}$ wyznaczyliśmy już we wzorze (Z1.37). Modyfikując więc (Z1.39) zgodnie z powyższymi informacjami, otrzymujemy ostateczny wzór na drugą pochodną błędu sieci neuronowej względem pierwszej wagi w warstwie wyjściowej, drugiej natomiast w warstwie ukrytej:

$$\begin{aligned} \frac{\partial^2 E_k}{\partial w_{ij}^{(1)} \partial w_{1t}^{(2)}} &= -y_j^{(0)} \delta_1^{(2)} \varphi'(net_t^{(1)}) \Delta_{ti} + y_j^{(0)} y_t^{(1)} \mathbf{H}_1 w_{1i}^{(2)} \varphi'(net_i^{(1)}) = \\ &= -y_j^{(0)} \left(\delta_1^{(2)} \varphi'(net_t^{(1)}) \Delta_{ti} - y_t^{(1)} w_{1i}^{(2)} \varphi'(net_i^{(1)}) \mathbf{H}_1 \right) \end{aligned} \tag{Z1.40}$$

Podobnie jak w poprzednim punkcie, $\delta_1^{(2)}$ jest standardowym błędem neuronu wyjściowego z metody wstecznej propagacji błędu (wzór (Z1.10)). Wartość $\mathbf{H}_1$ wyznaczamy ze wzoru (Z1.31b).

Wzory (Z1.34), (Z1.38) i (Z1.40) pozwalają, przy symetrycznym charakterze macierzy hesjanu, obliczyć wszystkie jej elementy. Pamiętajmy, że musimy je zastosować dla każdego wzorca treningowego sieci, a następnie zsumować razem wyniki otrzymane dla każdej z pochodnych.

Zauważmy jeszcze, że w każdym z tych wzorów występuje wielkość $\mathbf{H}_1$, którą musimy wyznaczyć dla neuronu wyjściowego za pomocą ogólnej zależności (Z1.31b). Dla błędu kwadratowego sieci i jego komponentów E_k, związanych z poszczególnymi obserwacjami treningowymi (Z1.3), zachodzą pewne poważne uproszczenia. Pamiętajmy bowiem, że w tym przypadku pierwsza i druga pochodna błędu względem stanu neuronu wyjściowego sieci wynoszą odpowiednio:

$$\begin{aligned} \frac{\partial E_k}{\partial y_1^{(wy)}} &= \frac{\partial}{\partial y_1^{(wy)}} \left(\frac{1}{2} \left(y_k - y_1^{(wy)} \right)^2 \right) = \left(y_k - y_1^{(wy)} \right) \\ \frac{\partial^2 E_k}{\partial y_1^{(wy)} \partial y_1^{(wy)}} &= -1 \end{aligned} \tag{Z1.41}$$

gdzie y_k jest obserwowaną wartością zmiennej zależnej (wyjściem treningowym) dla k-tej obserwacji w zbiorze uczącym sieci. Pamiętajmy również, że stan neuronu wyjściowego $y_1^{(wy)}$ wyznaczany jest dla wzorca wejściowego tej obserwacji $\mathbf{x}_k$.

Biorąc pod uwagę (Z1.41), zależność (Z1.31b), która pozwala na wyznaczenie wartości $\mathbf{H}_1$, możemy zapisać:

$$\mathbf{H}_1 = \varphi''(net_1^{(wy)})\left(y_k - y_1^{(wy)}\right) - \left(\varphi'(net_1^{(wy)})\right)^2 \tag{Z1.42}$$

Zauważmy jeszcze, że w przypadku logistycznej funkcji aktywacji neuronów $\varphi(net)$, jaką wykorzystywaliśmy w naszych pracach, istnieją proste procedury obliczania pierwszej i drugiej pochodnej za pomocą wartości tej funkcji (stanów wyjściowych neuronów), co pozwala jeszcze bardziej ułatwić implementację algorytmu wyznaczania hesjanu:

$$\begin{aligned}\varphi'(net) &= \left(\frac{1}{1+e^{-net}}\right)' = \frac{e^{-net}}{(1+e^{-net})^2} = \\ &= \frac{1}{(1+e^{-net})}\frac{1+e^{-net}-1}{(1+e^{-net})} = \frac{1}{(1+e^{-net})}\left(\frac{1+e^{-net}}{(1+e^{-net})} - \frac{1}{(1+e^{-net})}\right) = \\ &= \varphi(net)(1-\varphi(net))\end{aligned} \tag{Z1.43}$$

$$\begin{aligned}\varphi''(net) &= \left(\varphi(net)(1-\varphi(net))\right)' = \\ &= \varphi'(net)(1-\varphi(net)) - \varphi'(net)\varphi(net) = \\ &= \varphi'(net)(1-\varphi(net)-\varphi(net)) = \\ &= \varphi(net)(1-\varphi(net))(1-2\varphi(net))\end{aligned} \tag{Z1.44}$$

Z1.3. Wyznaczanie pochodnych wyjścia sieci MLP względem wag, dla danego wejścia

Podobnie jak w poprzednich punktach przyjmujemy strukturę sieci MLP z jedną warstwą ukrytą, postaci (Z1.1) lub (Z1.2), ale przedstawimy również krótko wyniki dla ogólniejszej architektury o dowolnej liczbie warstw ukrytych. Zakładamy jednak regresyjny charakter modelu, którego wyjście ma charakter skalarny, mówimy więc o wyznaczaniu jego gradientu względem współczynników wagowych. W przypadku sieci posiadających wiele wyjść wystarczy prezentowane dalej rozważania powtórzyć oddzielnie dla każdego z nich.

Gradient sieci względem współczynników wagowych dla danego wzorca wejściowego $\mathbf{x} = (x_1, \ldots, x_n)$ wykorzystywany jest w rozdziale 3.3.3 do oszacowania wariancji wyjściowej modelu (prognozy) dla tego wejścia. Gradienty sieci dla poszczególnych wzorców treningowych stosowane są także przy aproksymacji iloczynem skalarnym (Levenberga–Marquarda) macierzy hesjanu błędu modelu (również patrz rozdział 3.3.3). Ponadto pochodne sieci względem wag

obliczane są oczywiście w sposób niejawny podczas wyznaczania gradientu błędu w algorytmie wstecznej propagacji i w innych metodach uczenia sieci.

Z tego ostatniego powodu sposób wyznaczania gradientu wyjścia sieci stanowi prostą modyfikację algorytmu przedstawionego w punkcie Z1.1. W związku z tym zaprezentujemy jedynie zarys wyprowadzenia odpowiednich zależności, odnośnie do szczegółów odsyłając do odpowiednich obliczeń w powyższym punkcie.

Również i w tym przypadku oprzemy się na prostej obserwacji, że odwzorowanie realizowane przez każdy neuron jest złożeniem pobudzenia neuronu oraz funkcji aktywacji. Pozwala nam to na zastosowanie łańcuchowych reguł opartych na różniczkowaniu funkcji złożonej, postaci:

$$\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial w_{ij}^{(m)}} = \frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_i^{(m)}} \frac{\partial net_i^{(m)}}{\partial w_{ij}^{(m)}} \tag{Z1.45}$$

Po zastosowaniu obliczeń analogicznych do (Z1.6)–(Z1.8), otrzymujemy ogólny wzór na pochodną wyjścia sieci względem dowolnej wagi:

$$\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial w_{ij}^{(m)}} = -d_i^{(m)} y_j^{(m-1)} \tag{Z1.46}$$

gdzie przez $d_i^{(m)}$ oznaczyliśmy pochodną wyjścia sieci względem pobudzenia neuronu, do którego przypisana jest waga $w_{ij}^{(m)}$, pomnożoną przez –1:

$$d_i^{(m)} = -\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_i^{(m)}} \tag{Z1.47}$$

Konsekwentnie korzystając dalej z rozważań w punkcie Z1.1, zastępując w zasadzie jedynie pochodną błędu sieci przez pochodną jej wyjścia, otrzymujemy rekurencyjną regułę wstecznej propagacji dla współczynników $d_i^{(m)}$. Wartość początkowa dla neuronu wyjściowego przyjmuje następującą postać:

$$d_1^{(wy)} = -\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_1^{(wy)}} = \frac{\partial y_1^{(wy)}}{\partial net_1^{(wy)}} = -\varphi'(net_1^{(wy)}) \tag{Z1.48}$$

Dla sieci MLP o większej liczbie warstw ukrytych, o h_m neuronów w każdej warstwie łańcuchowa formuła wyznaczania współczynników $d_i^{(m)}$ w warstwie m, przy wykorzystaniu ich wartości $d_i^{(m+1)}$ w kolejnej warstwie, ma dokładnie taką

samą postać jak odpowiednia zależność dla błędów $\delta_i^{(m)}$ (patrz wzory (Z1.13) i (Z1.14)):

$$\begin{aligned} d_i^{(m)} &= -\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_i^{(m)}} = -\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial y_i^{(m)}} \varphi'(net_i^{(m)}) = \\ &= -\varphi'(net_i^{(m)}) \sum_{t=1}^{h_{m+1}} \frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_t^{(m+1)}} \frac{\partial net_t^{(m+1)}}{\partial y_i^{(m)}} = \\ &= \varphi'(net_i^{(m)}) \sum_{t=1}^{h_{m+1}} d_t^{(m+1)} w_{ti}^{(m+1)} \end{aligned} \tag{Z1.49}$$

W przypadku warstwowej sieci perceptronowej z jedną warstwą ukrytą i jednym neuronem wyjściowym danej przy użyciu zależności (Z1.1) lub (Z1.2) przedstawiona wyżej formuła wstecznej propagacji wielkości $d_i^{(m)}$ redukuje się do następujących prostych wzorów:

$$d_1^{(2)} = -\varphi'(net_1^{(2)}) \tag{Z1.50a}$$

$$d_i^{(1)} = d_1^{(2)} w_{1i}^{(2)} \varphi'(net_i^{(1)}) \text{ , } i = 1, \ldots, h \tag{Z1.50b}$$

Podsumowując, stwierdzamy, że w przypadku warstwowych sieci perceptronowych MLP, rozważanych w rozdziale 2 do prognozy krótkoterminowego zapotrzebowania na energię, proces wyznaczania gradientu wyjścia sieci neuronowej względem wag, przy danym wzorcu wejściowym $\mathbf{x} = (x_1, \ldots, x_n)$, przyjmuje postać:

1. Podajemy wzorzec **x** na wejście sieci. Przy wykorzystaniu zależności (3.6.1) (lub (Z1.2)) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Znajdujemy za pomocą (Z1.50a) współczynnik $d_1^{(2)}$ dla neuronu warstwy wyjściowej, a następnie za pomocą (Z1.50b) współczynniki $d_i^{(1)}$ dla wszystkich neuronów warstwy ukrytej.

3. Korzystając z (Z1.46) i obliczonych wartości d, wyznaczamy pochodne modelu względem wszystkich wag sieci, pamiętając oczywiście, że $y_i^{(0)} = x_i$.

Z1.4. Wyznaczanie pochodnych wyjścia sieci MLP względem zmiennych wejściowych

I znowu, jak w poprzednich punktach, dokładny algorytm wyznaczania jakobianu przedstawimy dla sieci MLP z jedną warstwą ukrytą, postaci (Z1.1) lub (Z1.2), ale zaprezentujemy również jego zarys dla ogólniejszej architektury o dowolnej liczbie warstw ukrytych. Również jak w poprzednich punktach ograniczamy się jednak do modeli o charakterze regresyjnym, o jednym neuro-

nie wyjściowym. Także i obecnie będziemy więc mówić o wyznaczaniu wektora pochodnych, czyli gradientu wyjścia sieci, względem poszczególnych zmiennych wejściowych $\mathbf{x} = (x_1, \ldots, x_n)$. W przypadku modelu o większej liczbie neuronów w warstwie wyjściowej aby wyznaczyć macierz jakobianu, wystarczy prezentowany w tym punkcie algorytm powtórzyć oddzielnie dla każdego wyjścia.

Pochodną wyjścia sieci MLP względem zmiennych wejściowych wykorzystujemy w punkcie 3.5 do propagacji błędów wejściowych i szacowania związanej z nimi niepewności prognozy. Używane są one również w wielu zagadnieniach pozostających poza zakresem tematycznym naszej pracy, takich jak analiza wrażliwości wejść, dobór optymalnej struktury sieci itp. Kolejnym ważnym obszarem zastosowań, w którym wymaga się często wyznaczenia gradientów sieci, są problemy optymalizacji w przestrzeni wejść, w których predyktor neuronowy stanowi model minimalizowanego (maksymalizowanego) odwzorowania.

Algorytm wyznaczania pochodnych wyjścia warstwowej sieci perceptronowej MLP względem wejść jest, w zasadzie, lustrzanym odbiciem algorytmu obliczania pochodnych względem wag z poprzedniego punktu Z1.3. Realizuje się go za pomocą analogicznej rekurencyjnej procedury wstecznej propagacji wartości pochodnej względem pobudzenia neuronów, począwszy od neuronu wyjściowego, następnie przez warstwę ukrytą (kolejne warstwy ukryte), w stronę wejściowej. Pamiętamy, że stany neuronów warstwy wejściowej reprezentują poszczególne wejścia sieci. Dokładniej mówiąc, jak pokażemy za chwilę, sama procedura wstecznej propagacji jest identyczna jak w przypadku gradientu sieci w przestrzeni parametrów. Różnica polega jedynie na finalnym obliczeniu pochodnej.

Rozważmy najpierw przypadek ogólniejszy – sieci MLP o jednym neuronie wyjściowym, ale dowolnej liczbie warstw ukrytych, o h_m neuronów w każdej warstwie. Zauważmy, że aby obliczyć pochodną wyjścia sieci względem danej zmiennej wejściowej x_i (stanu neuronu wejściowego $y_i^{(0)}$), możemy uwzględnić fakt, że $y_i^{(0)}$ stanowi wejście każdego neuronu z pierwszej warstwy ukrytej. W związku z tym pochodną tę możemy przedstawić następująco:

$$\begin{aligned}\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial x_i} &= \frac{\partial f(\mathbf{x},\mathbf{w})}{\partial y_i^{(0)}} = \\ &= \sum_{t=1}^{h_1} \frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_t^{(1)}} \frac{\partial net_t^{(1)}}{\partial y_i^{(0)}} = \sum_{t=1}^{h_1} \frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_t^{(1)}} w_{ti}^{(1)}\end{aligned} \tag{Z1.51}$$

Zauważmy, że pochodną wyjścia sieci względem pobudzenia możemy zastąpić przez zdefiniowaną w poprzednim punkcie, za pomocą wzoru (Z1.47), wartość $d_i^{(1)}$. Otrzymujemy więc natychmiast:

$$\frac{\partial f(\mathbf{x},\mathbf{w})}{\partial x_i} = \sum_{t=1}^{h_1} \frac{\partial f(\mathbf{x},\mathbf{w})}{\partial net_t^{(1)}} w_{ti}^{(1)} = -\sum_{t=1}^{h_1} d_t^{(1)} w_{ti}^{(1)} \tag{Z1.52}$$

Do obliczenia wartości $d_i^{(1)}$ możemy wykorzystać dokładnie łańcuchową metodę wstecznej propagacji, zdefiniowaną przez wzory (Z1.48) i (Z1.49), która w przypadku sieci MLP z jedną warstwą ukrytą upraszcza się do formuły danej przez (Z1.50). W przypadku sieci neuronowej o kilku wyjściach powyższa procedura dawałaby jeden wiersz macierzy jakobianu (dla jednej składowej funkcji), więc podobne kroki należałoby powtórzyć dla każdego neuronu wyjściowego.

Ograniczając się więc do przypadku wykorzystywanych przez nas w rozdziale 2 warstwowych sieci perceptronowych MLP, z jedną warstwą ukrytą, określonych przez zależność (Z1.1) lub (Z1.2), proces wyznaczania gradientu wyjścia sieci neuronowej względem poszczególnych zmiennych wejściowych, w punkcie $\mathbf{x} = (x_1, \ldots, x_n)$, możemy przedstawić następująco:

1. Podajemy wzorzec **x** na wejście sieci. Przy wykorzystaniu zależności (3.6.1) (lub (Z1.2)) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Znajdujemy za pomocą (Z1.50a) współczynnik $d_1^{(2)}$ dla neuronu warstwy wyjściowej, a następnie za pomocą (Z1.50b) – współczynniki $d_i^{(1)}$ dla wszystkich neuronów warstwy ukrytej.

3. Na podstawie obliczonych wartości $d_i^{(1)}$, na podstawie (Z1.52), wyznaczamy pochodne modelu względem poszczególnych wejść.

ZAŁĄCZNIK 2

Ważniejsze gradienty i hesjany związane z siecią neuronowo-rozmytą FBF

Z2.1. Wyznaczanie gradientu błędu sieci FBF względem wag, dla danego wzorca treningowego

Model sieci FBF (*Fuzzy Basis Function*) wykorzystywaliśmy do prognoz zapotrzebowania na energię elektryczną omawianych w punkcie 2.3.2. Przypomnijmy, że jest to addytywny system neuronowo-rozmyty, w którym wykorzystuje się iloczynową regułę wnioskowania Larsena oraz uproszczoną metodę wnioskowania rozmytego. Funkcje przynależności zbiorów rozmytych A_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$ wykorzystywanych w poprzednikach poszczególnych reguł systemu mają charakter funkcji Gaussa. Strukturę sieci możemy przedstawić przy użyciu następującego równania:

$$y(x_1,...,x_n) = \frac{\sum_{i=1}^{K} b_i^* \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^2}\right)}{\sum_{i=1}^{K} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^2}\right)} \qquad \text{(Z2.1a)}$$

Współczynniki (wagi) sieci neuronowo-rozmytej, a_{ij}^* i σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$, są parametrami krzywej Gaussa i mogą być interpretowane jako środki i szerokości funkcji przynależności zbiorów rozmytych poprzedników reguł A_{ij}, natomiast b_i^*, $i = 1, \ldots, K$, są środkami zbiorów rozmytych B_i występujących w następnikach reguł systemu.

Zauważmy, że równanie modelu FBF alternatywnie przedstawić możemy (rysunek 2.3.1 w rozdziale 2.3.2) w postaci sieci neuronowej o trzech warstwach węzłów:

$$\tau_i = \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{x_j - a_{ij}^*}{\sigma_{ij}}\right)^2\right), i = 1,...,K$$
$$v_i = \frac{\tau_i}{\sum_{t=1}^{K}\tau_t} \qquad \text{(Z2.1b)}$$
$$y = \sum_{i=1}^{K} v_i b_i^*$$

Gradient błędu kwadratowego sieci FBF względem wag wykorzystywany jest przede wszystkim w algorytmie jej uczenia metodą wstecznej propagacji błędu (lub innymi metodami gradientowymi), więc, podobnie jak w przypadku sieci MLP w punkcie Z1.1, pokażemy sposób jego wyznaczania dla danego wzorca treningowego.

Przyjmijmy więc, że zbiór $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$ stanowi zbiór danych treningowych, dla których wyznaczono błąd E. Zakładamy oczywiście kwadratową postać funkcji błędu, daną przez (Z1.3). Podobnie jak w punkcie Z1.1, jeżeli operację różniczkowania błędu wykonujemy dla danego wejściowego wzorca treningowego $\mathbf{x}_k$, to zauważmy, że występujące w funkcji błędu wyjścia modelu są stałe, z wyjątkiem jedynego wyjścia zależącego od $\mathbf{x}_k$, czyli $y(\mathbf{x}_k)$. Ich pochodna jest więc równa zero. W konsekwencji, dla danego $\mathbf{x}_k$, gradient błędu E redukuje się tylko do gradientu składnika błędu E_k odpowiadającego tej obserwacji treningowej:

$$E_k = \frac{1}{2}(y_k - y(x_{k1},...,x_{kn}))^2 = \frac{1}{2}(y_k - y(\mathbf{x}_k))^2 \qquad \text{(Z2.2)}$$

Analogicznie jak w punkcie Z1.1, w dalszej części naszych rozważań dla wszystkich pobudzeń i stanów neuronów sieci FBF pomijać będziemy dla uproszczenia indeks k, pamiętając jednakże, że wartości te są wyznaczone dla konkretnego wejściowego wzorca treningowego $\mathbf{x}_k$.

Ponieważ neuron wyjściowy ma charakter liniowy, to wyznaczenie pochodnych błędu sieci FBF względem parametrów b_i^*, $i = 1, \ldots, K$ stanowi w zasadzie czystą formalność:

$$\frac{\partial E}{\partial b_i^*} = \frac{\partial E_k}{\partial b_i^*} = \frac{\partial(\frac{1}{2}(y_k - y)^2)}{\partial b_i^*} =$$
$$= -(y_k - y)\frac{\partial y}{\partial b_i^*} = -(y_k - y)\frac{\partial}{\partial b_i^*}\left(\sum v_i b_i^*\right) = -(y_k - y)v_i \qquad \text{(Z2.3)}$$

Nieco więcej czasu musimy poświęcić na znalezienie formuły na pochodne błędu względem parametrów a_{ij}^* oraz σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$. Zróbmy to po kolei, począwszy od neuronu wyjściowego, wstecz przez kolejne warstwy. Niech w_{ij} oznacza dowolną wagę a_{ij}^* lub σ_{ij} w pierwszej warstwie ukrytej sieci (zbiorach rozmytych poprzedników reguł). Zauważmy wówczas, że możemy zapisać:

$$\frac{\partial E}{\partial w_{ij}} = -(y_k - y)\frac{\partial y}{\partial w_{ij}} \tag{Z2.4}$$

Dalej, różniczkując z kolei wyjście sieci FBF względem dowolnej wagi w_{ij}, otrzymujemy następującą zależność, w której wykorzystuje się pochodne stanów neuronów drugiej warstwy ukrytej względem wag:

$$\frac{\partial y}{\partial w_{ij}} = \sum_{p=1}^{K} b_p^* \frac{\partial v_p}{\partial w_{ij}} \tag{Z2.5}$$

Zauważmy, że ze wzoru na różniczkowanie ilorazu funkcji możemy w następujący sposób wyznaczyć pochodną wyjścia neuronów drugiej warstwy ukrytej v_p względem parametrów wagowych w_{ij}:

$$\begin{aligned}\frac{\partial v_p}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}}\left(\frac{\tau_p}{\sum_{t=1}^{K}\tau_t}\right) = \\ &= \frac{\frac{\partial \tau_p}{\partial w_{ij}}\sum_{t=1}^{K}\tau_t - \frac{\partial}{\partial w_{ij}}\left(\sum_{t=1}^{K}\tau_t\right)\tau_p}{\left(\sum_{t=1}^{K}\tau_t\right)^2} = \frac{\frac{\partial \tau_p}{\partial w_{ij}}\sum_{t=1}^{K}\tau_t - \sum_{t=1}^{K}\frac{\partial \tau_t}{\partial w_{ij}}\tau_p}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\end{aligned} \tag{Z2.6}$$

Podstawiając otrzymaną zależność (Z2.6) do (Z2.5), otrzymujemy:

$$\frac{\partial y}{\partial w_{ij}} = \sum_{p=1}^{k} b_p^* \frac{\frac{\partial \tau_p}{\partial w_{ij}} \sum_{t=1}^{K} \tau_t - \sum_{t=1}^{K} \frac{\partial \tau_t}{\partial w_{ij}} \tau_p}{\left(\sum_{t=1}^{K} \tau_t \right)^2} =$$

$$= \frac{\sum_{p=1}^{K} b_p^* \frac{\partial \tau_p}{\partial w_{ij}} \sum_{t=1}^{K} \tau_t - \sum_{p=1}^{K} b_p^* \tau_p \sum_{t=1}^{K} \frac{\partial \tau_t}{\partial w_{ij}}}{\left(\sum_{t=1}^{K} \tau_t \right)^2} \tag{Z2.7}$$

Zauważmy dalej, że stan (wyjście) neuronów pierwszej warstwy ukrytej τ_p w większości przypadków, tzn. dla $p \neq i$ nie zależy od w_{ij}, w związku z tym odpowiednie pochodne będą miały wówczas wartość 0. Uwzględniając powyższy fakt w (Z2.7), możemy zależność tę zapisać następująco:

$$\frac{\partial y}{\partial w_{ij}} = \frac{b_i^* \frac{\partial \tau_i}{\partial w_{ij}} \sum_{t=1}^{K} \tau_t - \frac{\partial \tau_i}{\partial w_{ij}} \sum_{p=1}^{K} b_p^* \tau_p}{\left(\sum_{t=1}^{K} \tau_t \right)^2} = \frac{\partial \tau_i}{\partial w_{ij}} \frac{b_i^* \sum_{t=1}^{K} \tau_t - \sum_{p=1}^{K} b_p^* \tau_p}{\left(\sum_{t=1}^{K} \tau_t \right)\left(\sum_{t=1}^{K} \tau_t \right)} =$$

$$= \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \left(\frac{b_i^* \sum_{t=1}^{K} \tau_t}{\left(\sum_{t=1}^{K} \tau_t \right)} - \sum_{p=1}^{K} \frac{\tau_p}{\left(\sum_{t=1}^{K} \tau_t \right)} b_p^* \right) = \tag{Z2.8}$$

$$= \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \left(b_i^* - \sum_{p=1}^{K} \nu_p b_p^* \right) = \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \left(b_i^* - y \right)$$

Dla obu rodzajów wag neuronów pierwszej warstwy ukrytej sieci FBF, zarówno centrów funkcji przynależności zbiorów rozmytych poprzedników a_{ij}^*, jak i ich szerokości σ_{ij}, wzór (Z2.8) daje więc niemal jednolitą formułę obliczania pochodnej wyjścia sieci. Obydwa warianty różnić się będą między sobą

jedynie sposobem wyznaczania pochodnej wyjścia neuronu τ_i względem wagi danego typu. W przypadku stosowanej w sieciach FBF gaussowskiej funkcji przynależności (patrz (Z2.1b)), dla środków a_{ij}^*, otrzymujemy następującą zależność:

$$\begin{aligned}\frac{\partial \tau_i}{\partial a_{ij}^*} &= \frac{\partial}{\partial a_{ij}^*} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \left(\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}} \right)^2 \right) = \\ &= \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \left(\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}} \right)^2 \right) = \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \tau_i\end{aligned} \tag{Z2.9}$$

Korzystając więc z (Z2.4), (Z2.8) i (Z2.9), ostatecznie formułę na pochodną błędu kwadratowego sieci FBF względem parametrów środków zbiorów rozmytych poprzedników a_{ij}^*, dla danej obserwacji treningowej $\mathbf{x}_k$, możemy zapisać:

$$\begin{aligned}\frac{\partial E}{\partial a_{ij}^*} &= -(y_k - y) \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \tau_i \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \left(b_i^* - y \right) = \\ &= -(y_k - y) \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \nu_i \left(b_i^* - y \right)\end{aligned} \tag{Z2.10}$$

Analogicznie pochodną wyjścia neuronu τ_i względem parametru szerokości funkcji Gaussa σ_{ij} możemy obliczyć w następujący sposób:

$$\begin{aligned}\frac{\partial \tau_i}{\partial \sigma_{ij}} &= \frac{\partial}{\partial \sigma_{ij}} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \left(\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}} \right)^2 \right) = \\ &= \frac{(x_{kj} - a_{ij}^*)}{\sigma_{ij}^2} \frac{(x_{kj} - a_{ij}^*)}{\sigma_{ij}} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \left(\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}} \right)^2 \right) = \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} \tau_i\end{aligned} \tag{Z2.11}$$

Podobnie jak dla a_{ij}^*, również korzystając z wcześniej obliczonych zależności (Z2.4), (Z2.8) i oczywiście w tym przypadku z (Z2.11), otrzymujemy ostateczną formułę dla pochodnej błędu kwadratowego sieci FBF względem parametrów szerokości zbiorów rozmytych poprzedników reguł systemu σ_{ij}, dla danej obserwacji treningowej $\mathbf{x}_k$:

$$\frac{\partial E}{\partial \sigma_{ij}} = -(y_k - y)\frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3}\tau_i \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) = \\ = -(y_k - y)\frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3}\nu_i\left(b_i^* - y\right) \tag{Z2.12}$$

Podsumowując nasze rozważania, przedstawiony w tym punkcie proces wyznaczania gradientu błędu sieci FBF dla danej obserwacji treningowej $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$ przyjmuje postać następującego algorytmu:

1. Podajemy wzorzec $\mathbf{x}_k$ na wejście sieci FBF; przy wykorzystaniu zależności (3.7.1a) lub (Z2.1b) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Za pomocą zależności (Z2.3) obliczamy pochodne błędu kwadratowego sieci względem wag neuronu wyjściowego b_i^*, $i = 1, \ldots, K$.

3. Za pomocą zależności (Z2.10) obliczamy pochodne błędu kwadratowego względem parametrów środków krzywej Gaussa w pierwszej warstwie ukrytej sieci, a_{ij}^*, $j = 1,\ldots, n$, $i = 1, \ldots, K$.

4. Za pomocą zależności (Z2.12) obliczamy pochodne błędu kwadratowego względem parametrów szerokości krzywej Gaussa w pierwszej warstwie ukrytej sieci, σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$.

Przy bezpośrednim wykorzystaniu otrzymanych zależności dla poszczególnych pochodnych (Z2.3), (Z2.10) i (Z2.12) kolejność kroków 2–4 algorytmu wyznaczania gradientu błędu jest w zasadzie dowolna. Porównując jednak powyższe wzory, łatwo można zauważyć, że układają się one w pewną formułę łańcuchową, której zastosowanie warto rozważyć ze względów implementacyjnych:

$$\frac{\partial E}{\partial b_i^*} = -(y_k - y)\nu_i \tag{Z2.13a}$$

$$\frac{\partial E}{\partial a_{ij}^*} = -(y_k - y)\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}\nu_i\left(b_i^* - y\right) = \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}\left(b_i^* - y\right)\frac{\partial E}{\partial b_i^*} \tag{Z2.13b}$$

$$\frac{\partial E}{\partial \sigma_{ij}} = -(y_k - y)\frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3}\nu_i\left(b_i^* - y\right) = \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}}\frac{\partial E}{\partial a_{ij}^*} \tag{Z2.13c}$$

Z2.2. Wyznaczanie hesjanu błędu kwadratowego sieci FBF względem wag

Zastanówmy się obecnie nad problemem wyznaczenia macierzy **H** drugich pochodnych (hesjanu) funkcji błędu sieci neuronowo-rozmytej FBF względem wag. Przypomnijmy, że odwrotność hesjanu błędu wykorzystujemy w przedsta-

wionej w punkcie 3.3.2 metodzie delta do oszacowania macierzy kowariancji parametrów modelu i obliczenia odchylenia standardowego (lub wariancji) warunkowego rozkładu wartości wyjścia modelu dla danego wzorca wejściowego (rozkładu otrzymywanej prognozy). Jak wspomnieliśmy w punkcie Z1.2, w przypadku warstwowej sieci perceptronowej MLP hesjan błędu może być również potrzebny do niektórych bardziej zaawansowanych algorytmów uczenia (bądź douczania) sieci FBF.

W bieżącym punkcie interesować nas będzie wyłącznie metoda dokładnego wyznaczania hesjanu błędu kwadratowego, zbudowana konkretnie dla struktury sieci neuronowo-rozmytej FBF. Przypomnijmy tylko, że istnieje szereg metod przybliżonego szacowania macierzy **H**, które mają charakter uniwersalny i mogą być stosowane (przynajmniej do pewnego miejsca) dla modeli różnego typu. W interesującym nas kontekście wyznaczania macierzy kowariancji parametrów modelu najważniejszą z nich jest metoda aproksymacji iloczynem skalarnym (nazywana również aproksymacją Levenberga–Marquarda), którą przedstawiliśmy w punkcie 3.3.2.

Oznaczenia stosowane w bieżącym punkcie są takie same jak w punkcie poprzednim, Z2.1, chyba że zostanie wyraźnie stwierdzone inaczej. Nasz cel polega na znalezieniu drugich pochodnych dla dowolnej pary parametrów. Pamiętajmy jednak, że macierz hesjanu ma charakter symetryczny, więc założenie jakiegoś uporządkowania kolejności parametrów, według których wyznaczane są pochodne, np. związanego z ich położeniem w strukturze sieci FBF, nie zmniejsza ogólności rozważań. Dla układu wag w odwrotnej kolejności możemy skorzystać właśnie z symetryczności macierzy **H**.

Przyjmijmy, tak samo jak w poprzednim punkcie, że sieć dopasowana została na zbiorze danych $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$, a więc różniczkowany błąd kwadratowy E (Z1.3) wyznaczono na tym właśnie zbiorze. Ponownie rozważmy błąd E na całym zbiorze treningowym, jako sumę błędów E_k określonych przez (Z2.2), związanych z odchyleniami poszczególnych obserwacji treningowych. Podobnie jak w punkcie Z1.2, dla sieci MLP elementy hesjanu błędu E sieci FBF będziemy więc wyznaczać odrębnie, obliczając drugie pochodne poszczególnych składników błędu E_k, a następnie sumując je po wszystkich wzorcach treningowych.

Zajmijmy się teraz oszacowaniem drugiej pochodnej składnika pojedynczego błędu E_k. Tym razem również, tak jak w poprzednim punkcie, indeks k dla wszystkich pobudzeń i stanów neuronów będziemy dla uproszczenia pomijać, pamiętając jednakże, że ich wartości są wyznaczone dla konkretnego wejściowego wzorca treningowego $\mathbf{x}_k$.

Zdecydowanie najprostsze będzie wyznaczenie drugich pochodnych względem wag neuronu wyjściowego (parametrów b_i^*, $i = 1, \ldots, K$). Korzystając ze wzoru (3.7.3), niemal natychmiast możemy otrzymać następującą formułę:

$$\frac{\partial^2 E_k}{\partial b_j^* \partial b_i^*} = \frac{\partial\left(-(y_k - y)v_i\right)}{\partial b_j^*} = v_j v_i \qquad \text{(Z2.14)}$$

Wzór (Z2.14) nie powinien być zresztą specjalną niespodzianką. Neuron wyjściowy sieci ma charakterystykę liniową, z parametrami b_i^* i wejściami v_i. Pamiętając o związku między hesjanem błędu a macierzą kowariancji parametrów modelu (3.3.10), bez trudu możemy zauważyć, że zsumowane dla wszystkich wzorców treningowych pochodne (Z2.14) odpowiadają dokładnie macierzy kowariancji modelu liniowego, określonej przez zależność (3.2.31).

W kolejnym kroku wyznaczmy pochodne względem parametrów b_m^*, $m = 1,\ldots, n$ oraz a_{ij}^* i σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$. Oznaczmy, jak w poprzednim punkcie, przez w_{ij} dowolną wagę a_{ij}^* lub σ_{ij} w pierwszej warstwie ukrytej sieci neuronowo-rozmytej (zbiorach rozmytych poprzedników reguł). Ponieważ nie korzystamy tutaj z formuł łańcuchowych do obliczania pochodnych, więc w przeciwieństwie do algorytmu wyznaczania hesjanu błędu sieci perceptronowej MLP (patrz rozdział Z1.2), kolejność wyboru wag podczas różniczkowania funkcji błędu sieci FBF nie ma znaczenia. Nieco wygodniej będzie nam zróżniczkować E_k najpierw względem wagi z warstwy ukrytej w_{ij}, a następnie względem wagi neuronu wyjściowego b_m^*. Bez większego trudu można obliczyć, że kolejność odwrotna daje dokładnie ten sam wynik.

Na podstawie otrzymanych w poprzednim podrozdziale Z2.1 wzorów dla pochodnych (Z2.4) i (Z2.8) otrzymujemy następującą zależność dla pochodnej błędu E_k względem wagi warstwy ukrytej w_{ij}:

$$\frac{\partial E_k}{\partial w_{ij}} = -(y_k - y)\frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(b_i^* - y\right) \qquad \text{(Z2.15)}$$

Jeżeli zależność (Z2.15) zróżniczkujemy względem dowolnej wagi neuronu wyjściowego b_m^*, otrzymujemy następującą formułę dla odpowiedniej drugiej pochodnej błędu kwadratowego E_k:

$$\frac{\partial^2 E_k}{\partial b_m^* \partial w_{ij}} = \frac{\partial}{\partial b_m^*}\left(-(y_k - y)\frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(b_i^* - y\right)\right) =$$

$$= \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(\frac{\partial y}{\partial b_m^*}\left(b_i^* - y\right) - (y_k - y)\left(\frac{\partial b_i^*}{\partial b_m^*} - \frac{\partial y}{\partial b_m^*} \right)\right) = \tag{Z2.16}$$

$$= \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(v_m \left(b_i^* - y\right) - (y_k - y)(\Delta_{im} - v_m)\right)$$

gdzie Δ_{im} oznacza deltę Kroneckera, tzn. jest równe 1 dla $i = m$, 0 w przeciwnym przypadku.

Wykorzystując dalej wzory (Z2.9) i (Z2.11) na pochodne wyjścia neuronu warstwy ukrytej względem parametrów a_{ij}^* i σ_{ij}, możemy wyznaczyć odpowiednie elementy hesjanu błędu E_k. Na podstawie (Z2.9) dla pochodnej względem wag a_{ij}^* i b_m^* możemy niemal natychmiast napisać:

$$\frac{\partial^2 E_k}{\partial b_m^* \partial a_{ij}^*} = \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{\tau_i}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(v_m \left(b_i^* - y\right) - (y_k - y)(\Delta_{im} - v_m)\right) =$$

$$= \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} v_i \left(v_m \left(b_i^* - y\right) - (y_k - y)(\Delta_{im} - v_m)\right) \tag{Z2.17}$$

Podobnie korzystając z (Z2.11), otrzymujemy zależność dla drugiej pochodnej błędu E_k względem wag σ_{ij} i b_m^*:

$$\frac{\partial^2 E_k}{\partial b_m^* \partial \sigma_{ij}} = \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} \frac{\tau_i}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(v_m \left(b_i^* - y\right) - (y_k - y)(\Delta_{im} - v_m)\right) =$$

$$= \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} v_i \left(v_m \left(b_i^* - y\right) - (y_k - y)(\Delta_{im} - v_m)\right) \tag{Z2.18}$$

Gdy obie wagi, względem których wyznaczamy drugą pochodną błędu, pochodzą z warstwy ukrytej, obliczenia stają się może nieco bardziej żmudne, ale ich skala trudności specjalnie nie wzrasta. Niech więc w_{ij} i w_{ml} będą dowolnymi wagami sieci FBF, pochodzącymi z pierwszej warstwy ukrytej (przypomnijmy,

że druga warstwa ukryta nie ma parametrów adaptacyjnych). Chwilowo nie rozróżniamy, czy w_{ij} i w_{ml} są parametrami środków czy szerokości funkcji Gaussa definiującej funkcje przynależności zbiorów rozmytych poprzedników reguł systemu neuronowo-rozmytego.

Przypomnijmy sobie wzór (3.3.15) wykorzystywany podczas oszacowania hesjanu błędu kwadratowego modelu metodą aproksymacji iloczynem skalarnym (aproksymacji Levenberga–Marquarda). Ma on zastosowanie do błędu kwadratowego i uzależnia drugą pochodną tego błędu od pierwszych i drugich pochodnych wyjścia modelu. Jego wyprowadzenie jest w zasadzie elementarne i zostało już zaprezentowane w punkcie 3.3, więc nie będziemy go tutaj powtarzać – odsyłamy Czytelnika do odpowiedniego wcześniejszego fragmentu pracy. Pozwoli on nam na niewielkie uproszczenie i uporządkowanie obliczeń. Mianowicie na podstawie wzoru (3.3.15) możemy, dostosowując oznaczenia do obecnego przypadku, drugą pochodną funkcji błędu sieci zapisać w następujący sposób:

$$\frac{\partial^2 E_k}{\partial w_{ml}\partial w_{ij}} = \frac{\partial y}{\partial w_{ml}}\frac{\partial y}{\partial w_{ij}} - (y_k - y)\frac{\partial^2 y}{\partial w_{ml}\partial w_{ij}} \tag{Z2.19}$$

Oznaczmy pierwszą pochodną wyjścia sieci względem wagi w_{ij} przez y'_{ij}, natomiast jego drugą pochodną względem wag w_{ij} oraz w_{ml}, przez $y''_{ml,ij}$. Przy tych oznaczeniach możemy przepisać (Z2.19) następująco:

$$\frac{\partial^2 E_k}{\partial w_{ml}\partial w_{ij}} = y'_{ml}y'_{ij} - (y_k - y)y''_{ml,ij} \tag{Z2.20}$$

Zauważmy dalej, że interesuje nas obecnie element hesjanu błędu dla wag w_{ij} oraz w_{ml}, pochodzących z warstwy ukrytej, więc do obliczenia pochodnych wyjścia sieci względem tych wag, tj. wartości y'_{ij} oraz y'_{ml}, możemy skorzystać ze znalezionej w poprzednim punkcie Z2.1 formuły (Z2.8). Dla pierwszego członu wzoru (Z2.20) otrzymujemy więc następującą zależność:

$$\begin{aligned} y'_{ml}\,y'_{ij} &= \frac{\partial \tau_m}{\partial w_{ml}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_m^* - y\right)\frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) = \\ &= \frac{\partial \tau_m}{\partial w_{ml}}\frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\left(b_m^* - y\right)\left(b_i^* - y\right) \end{aligned} \tag{Z2.21}$$

W przypadku drugiej pochodnej wyjścia sieci $y''_{ml,ij}$ również możemy skorzystać z formuły (Z2.8), różniczkując powyższą zależność względem wagi w_{ml}. Dzięki temu, korzystając ze wzoru na pochodną iloczynu, otrzymujemy:

$$y''_{ml,ij} = \frac{\partial}{\partial w_{ml}}\left(\frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right)\right) = \frac{\partial^2 \tau_i}{\partial w_{ml}\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right)$$
$$- \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial \tau_m}{\partial w_{ml}}\frac{\partial \tau_i}{\partial w_{ij}}\left(b_i^* - y\right) - \frac{\partial y}{\partial w_{ml}}\frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)} \quad \text{(Z2.22)}$$

Obliczając pochodną w drugim członie (Z2.22), wykorzystaliśmy fakt, że stany (wyjścia) neuronów pierwszej warstwy ukrytej τ_t dla $t \neq m$ nie zależą od w_{ml} (ponieważ jest to wtedy waga innego neuronu), w związku z tym odpowiednie pochodne będą w takich przypadkach miały wartość 0. Zastępujemy jeszcze w ostatnim członie wzoru (Z2.8) pochodną wyjścia sieci względem wagi:

$$y''_{ml,ij} = \frac{\partial^2 \tau_i}{\partial w_{ml}\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) - \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial \tau_m}{\partial w_{ml}}\frac{\partial \tau_i}{\partial w_{ij}}\left(b_i^* - y\right)$$
$$- \frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_m^* - y\right)\frac{\partial \tau_m}{\partial w_{ml}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)} = \quad \text{(Z2.23)}$$
$$= \frac{\partial^2 \tau_i}{\partial w_{ml}\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) - \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial \tau_m}{\partial w_{ml}}\frac{\partial \tau_i}{\partial w_{ij}}\left(b_i^* + b_m^* - 2y\right)$$

Zauważmy dalej, że wyjście (stan) neuronu warstwy ukrytej τ_i nie zależy od wag innych neuronów, w związku z tym jego druga pochodna różni się od 0 jedynie, gdy jest liczona względem wag tego samego neuronu, tzn.:

$$\frac{\partial^2 \tau_i}{\partial w_{ml}\partial w_{ij}} = \Delta_{mi}\frac{\partial^2 \tau_i}{\partial w_{il}\partial w_{ij}} \quad \text{(Z2.24)}$$

gdzie Δ_{im} oznacza deltę Kroneckera, tzn. jest równe 1 dla $i = m$, 0 w przeciwnym przypadku.

Dla obydwu wag w_{ij} oraz w_{ml} pochodzących z różnych neuronów druga pochodna wyjścia sieci (Z2.23) upraszcza się do następującej zależności:

$$y''_{ml,ij} = -\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial\tau_m}{\partial w_{ml}}\frac{\partial\tau_i}{\partial w_{ij}}\left(b_i^* + b_m^* - 2y\right) \tag{Z2.25}$$

Podstawiając (Z2.21) i (Z2.25) do (Z2.20), otrzymujemy wzór na drugą pochodną błędu sieci neuronowo-rozmytej FBF względem wag pochodzących z różnych neuronów pierwszej warstwy ukrytej:

$$\begin{aligned}\frac{\partial^2 E_k}{\partial w_{ml}\partial w_{ij}} &= \frac{\partial\tau_m}{\partial w_{ml}}\frac{\partial\tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\left(b_m^* - y\right)\left(b_i^* - y\right) + \\ &+ (y_k - y)\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial\tau_m}{\partial w_{ml}}\frac{\partial\tau_i}{\partial w_{ij}}\left(b_i^* + b_m^* - 2y\right) = \\ &= \frac{\partial\tau_m}{\partial w_{ml}}\frac{\partial\tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\left(\left(b_m^* - y\right)\left(b_i^* - y\right) + (y_k - y)\left(b_i^* + b_m^* - 2y\right)\right)\end{aligned} \tag{Z2.26}$$

Przypomnijmy, że pochodne wyjścia neuronu warstwy ukrytej τ_i zarówno względem parametrów a_{ij}^*, jak i σ_{ij} dane są przez zależności (Z2.9) i (Z2.11). Wykorzystując powyższe wzory dla wag różnych neuronów, tj. dla $i \neq m$, otrzymujemy następujące formuły dla odpowiednich elementów hesjanu błędu:

– różniczkowanie względem a_{ij}^*, a następnie σ_{ml} (przypomnijmy, że w przypadku odwrotnej kolejności wag możemy dla obliczenia pochodnej skorzystać z symetryczności macierzy hesjanu):

$$\begin{aligned}\frac{\partial^2 E_k}{\partial\sigma_{ml}\partial a_{ij}^*} &= \frac{\partial\tau_m}{\partial\sigma_{ml}}\frac{\partial\tau_i}{\partial a_{ij}^*}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\left(\left(b_m^* - y\right)\left(b_i^* - y\right) + (y_k - y)\left(b_i^* + b_m^* - 2y\right)\right) = \\ &= \frac{(x_{kl} - a_{ml}^*)^2}{\sigma_{ml}^3}\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}\frac{\tau_m\tau_i}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\left(\left(b_m^* - y\right)\left(b_i^* - y\right) + (y_k - y)\left(b_i^* + b_m^* - 2y\right)\right) = \\ &= \frac{(x_{kl} - a_{ml}^*)^2}{\sigma_{ml}^3}\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}v_m v_i\left(\left(b_m^* - y\right)\left(b_i^* - y\right) + (y_k - y)\left(b_i^* + b_m^* - 2y\right)\right)\end{aligned} \tag{Z2.27a}$$

– różniczkowanie względem σ_{ij} oraz σ_{ml}:

$$\frac{\partial^2 E_k}{\partial\sigma_{ml}\partial\sigma_{ij}} = \\ = \frac{(x_{kl}-a_{ml}^*)^2}{\sigma_{ml}^3}\frac{(x_{kj}-a_{ij}^*)^2}{\sigma_{ij}^3} v_m v_i \left(\left(b_m^*-y\right)\left(b_i^*-y\right)+(y_k-y)\left(b_i^*+b_m^*-2y\right)\right) \tag{Z2.27b}$$

– różniczkowanie względem a_{ij}^* oraz a_{ml}^*:

$$\frac{\partial^2 E_k}{\partial a_{ml}^*\partial a_{ij}^*} = \\ = \frac{x_{kl}-a_{ml}^*}{\sigma_{ml}^2}\frac{x_{kj}-a_{ij}^*}{\sigma_{ij}^2} v_m v_i \left(\left(b_m^*-y\right)\left(b_i^*-y\right)+(y_k-y)\left(b_i^*+b_m^*-2y\right)\right) \tag{Z2.27c}$$

Wyznaczone powyżej wzory (Z2.27) pozwalają na znalezienie elementów drugich pochodnych błędu kwadratowego sieci FBF względem wag w_{ij} oraz w_{ml}, pochodzących z różnych neuronów warstwy ukrytej modelu. Rozważmy teraz odmienną sytuację, to jest przypadek, w którym obie wagi pochodzą z tego samego neuronu. Wówczas, po uwzględnieniu wzoru (Z2.24) oraz faktu że $i = m$, zależność (Z2.23) dla drugiej pochodnej wyjścia sieci względem obu tych wag upraszcza się do postaci:

$$y''_{il,ij} = \frac{\partial^2\tau_i}{\partial w_{il}\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^*-y\right) - \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial\tau_i}{\partial w_{il}}\frac{\partial\tau_i}{\partial w_{ij}}\left(b_i^*+b_i^*-2y\right) = \\ = \frac{\partial^2\tau_i}{\partial w_{il}\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^*-y\right) - 2\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial\tau_i}{\partial w_{il}}\frac{\partial\tau_i}{\partial w_{ij}}\left(b_i^*-y\right) = \tag{Z2.28}$$

$$= \left(b_i^*-y\right)\left(\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\frac{\partial^2\tau_i}{\partial w_{il}\partial w_{ij}} - 2\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\frac{\partial\tau_i}{\partial w_{il}}\frac{\partial\tau_i}{\partial w_{ij}}\right)$$

Podobnie możemy nieco uprościć wzór na iloczyn pierwszych pochodnych wyjścia sieci względem tych wag:

$$y'_{il}y'_{ij} = \frac{\partial \tau_i}{\partial w_{il}} \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \left(b_i^* - y\right)\left(b_i^* - y\right) = \\ = \frac{\partial \tau_i}{\partial w_{il}} \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \left(b_i^* - y\right)^2 \tag{Z2.29}$$

I znów, podstawiając (Z2.28) oraz (Z2.29) do ogólnej zależności na drugą pochodną błędu względem wag warstwy ukrytej, otrzymujemy:

$$\frac{\partial^2 E_k}{\partial w_{il} \partial w_{ij}} = \frac{\partial \tau_i}{\partial w_{il}} \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \left(b_i^* - y\right)^2 \\ -(y_k - y)\left(b_i^* - y\right)\left(\frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \frac{\partial^2 \tau_i}{\partial w_{il} \partial w_{ij}} - 2 \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \frac{\partial \tau_i}{\partial w_{il}} \frac{\partial \tau_i}{\partial w_{ij}} \right) \tag{Z2.30}$$

$$= \left(b_i^* - y\right)\left(\frac{\partial \tau_i}{\partial w_{il}} \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \left(b_i^* - 3y + 2y_k\right) - \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \frac{\partial^2 \tau_i}{\partial w_{il} \partial w_{ij}} (y_k - y) \right)$$

Ponownie skorzystamy z faktu, że pochodne wyjścia neuronu warstwy ukrytej τ_i zarówno względem parametrów a_{ij}^*, jak i σ_{ij} dane są przez zależności (Z2.9) i (Z2.11). Pozostaje więc nam jedynie obliczenie drugiej pochodnej. Po kolei więc wyprowadzimy wzory dla elementów hesjanu błędu kwadratowego sieci FBF, wynikające z różnych kombinacji a_{ij}^* i σ_{ij}.

Rozpocznijmy od drugiej pochodnej względem a_{ij}^*, a następnie σ_{il} (przypomnijmy, że również i obecnie, w przypadku odwrotnej kolejności wag, możemy do obliczenia pochodnej skorzystać z symetryczności macierzy hesjanu). Stosując (Z2.9) i (Z2.11), pochodną wyjścia neuronu warstwy ukrytej względem tych wag możemy zapisać następująco:

$$\frac{\partial^2 \tau_i}{\partial \sigma_{il} \partial a_{ij}^*} = \frac{\partial}{\partial \sigma_{il}} \left(\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \tau_i \right) = \tag{Z2.31}$$

$$= \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{\partial \tau_i}{\partial \sigma_{il}} = \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{(x_{kl} - a_{il}^*)^2}{\sigma_{il}^3} \tau_i$$

Podstawiając więc do (Z2.30) drugą pochodną wyjścia neuronu ze wzoru (Z2.31), zaś pierwsze pochodne z (Z2.9) i (Z2.11), otrzymujemy ostateczny wzór na drugą pochodną błędu kwadratowego sieci FBF, względem wag a_{ij}^*, a następnie σ_{il}, pochodzących z tego samego neuronu warstwy ukrytej:

$$\frac{\partial^2 E_k}{\partial \sigma_{il} \partial a_{ij}^*} = \left(b_i^* - y\right) \left(\frac{(x_{kl} - a_{il}^*)^2}{\sigma_{il}^3} \tau_i \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \tau_i \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)^2} \left(b_i^* - 3y + 2y_k\right) - \right.$$

$$\left. \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{(x_{kl} - a_{il}^*)^2}{\sigma_{il}^3} \tau_t (y_k - y) \right) = \tag{Z2.32}$$

$$= \left(b_i^* - y\right) \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{(x_{kl} - a_{il}^*)^2}{\sigma_{il}^3} v_i \left(v_i \left(b_i^* - 3y + 2y_k\right) - (y_k - y) \right)$$

W analogiczny sposób jak w przypadku zależności (Z2.31), różniczkując wyjście neuronu warstwy ukrytej najpierw względem σ_{ij}, a następnie względem σ_{il}, otrzymujemy drugą pochodną wyjścia neuronu warstwy ukrytej względem tych wag:

$$\frac{\partial^2 \tau_i}{\partial \sigma_{il} \partial \sigma_{ij}} = \frac{\partial}{\partial \sigma_{il}} \left(\frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ik}^3} \tau_i \right) = \tag{Z2.33}$$

$$= \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ik}^3} \frac{\partial \tau_i}{\partial \sigma_{il}} = \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ik}^3} \frac{(x_{kl} - a_{il}^*)^2}{\sigma_{il}^3} \tau_i$$

I dalej, podstawiając drugą pochodną wyjścia neuronu ze wzoru (Z2.33) i pierwsze pochodne z (Z2.9) i (Z2.11) do ogólnego wzoru (Z2.30), otrzymujemy ostateczną zależność na drugą pochodną błędu kwadratowego sieci FBF, względem wag σ_{ij}, a następnie σ_{il}, pochodzących z tego samego neuronu warstwy ukrytej:

$$\frac{\partial^2 E_k}{\partial\sigma_{il}\partial\sigma_{ij}} = \left(b_i^* - y\right)\left(\frac{(x_{kl}-a_{il}^*)^2}{\sigma_{il}^3}\tau_i \frac{(x_{kj}-a_{ij}^*)^2}{\sigma_{ij}^3}\tau_i \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\left(b_i^* - 3y + 2y_k\right) - \right. \tag{Z2.34}$$

$$\left. + \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)} \frac{(x_{kj}-a_{ij}^*)^2}{\sigma_{ik}^3} \frac{(x_{kl}-a_{il}^*)^2}{\sigma_{il}^3}\tau_i(y_k - y) \right) =$$

$$= \left(b_i^* - y\right)\frac{(x_{kj}-a_{ij}^*)^2}{\sigma_{ik}^3} \frac{(x_{kl}-a_{il}^*)^2}{\sigma_{il}^3} v_i\left(v_i\left(b_i^* - 3y + 2y_k\right) - (y_k - y)\right)$$

Do obliczenia pozostaje nam już ostatnia z potrzebnych drugich pochodnych wyjścia neuronu warstwy ukrytej, mianowicie względem wag a_{ij}^* i potem a_{il}^*. Wyznaczamy ją w niemal dokładnie taki sam sposób jak w poprzednich przypadkach, dla wzorów (Z2.31) lub (Z2.33). Niemal natychmiast otrzymujemy:

$$\frac{\partial^2 \tau_i}{\partial a_{il}^* \partial a_{ij}^*} = \frac{\partial}{\partial a_{il}^*}\left(\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}\tau_i\right) = \tag{Z2.35}$$

$$= \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}\frac{\partial\tau_i}{\partial a_{il}^*} = \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}\frac{x_{kl} - a_{il}^*}{\sigma_{il}^2}\tau_i$$

Ponownie podstawiamy drugą pochodną wyjścia neuronu ze wzoru (Z2.35) i pierwsze pochodne z (Z2.9) i (Z2.11) do ogólnego wzoru (Z2.30). Otrzymujemy w ten sposób ostatnią zależność dla drugiej pochodnej błędu kwadratowego sieci FBF, tym razem względem wag a_{ij}^* i potem a_{il}^*, pochodzących z tego samego neuronu warstwy ukrytej:

$$\frac{\partial^2 E_k}{\partial a_{il}^* \partial a_{ij}^*} = \left(b_i^* - y\right)\left(\frac{x_{kl} - a_{il}^*}{\sigma_{il}^2}\tau_i \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2}\tau_i \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\left(b_i^* - 3y + 2y_k\right) - \right.$$
$$\left. \frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)} \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{x_{kl} - a_{il}^*}{\sigma_{il}^2}\tau_i (y_k - y) \right) = \tag{Z2.36}$$
$$= \left(b_i^* - y\right)\frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{x_{kl} - a_{il}^*}{\sigma_{il}^2} v_i \left(v_i \left(b_i^* - 3y + 2y_k\right) - (y_k - y)\right)$$

Mamy już wszystkie składniki potrzebne do wyznaczenia elementów macierzy hesjanu błędu kwadratowego dla modelu sieci FBF. Algorytm nie jest może bardzo złożony koncepcyjnie, ale z powodu różnorodności typów wag występujących w modelu FBF wymaga zastosowania wielu różnych formuł do obliczenia poszczególnych bloków w macierzy **H**. Przedstawmy go więc w uporządkowanej postaci.

1. Obliczenia wykonujemy oddzielnie dla każdego wzorca danych występującego w zbiorze treningowym $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, określając wartości elementów hesjanu błędu E_k, przypadającego na ten wzór. Odpowiednie pochodne obliczone w kolejnych punktach algorytmu należy więc podsumować dla kolejnych obserwacji treningowych.

2. Podajemy wzorzec $\mathbf{x}_k$ na wejście sieci FBF. Przy wykorzystaniu zależności (3.7.1a) lub (Z2.1b) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

3. Wyznaczamy drugie pochodne błędu E_k w sytuacji, w której obie wagi względem których różniczkujemy, są wagami neuronu wyjściowego b_i^*, b_j^*. Do obliczenia pochodnej stosujemy wzór (Z2.14).

4. Wyznaczamy drugie pochodne błędu E_k w sytuacji, w której jedna z wag względem której różniczkujemy, jest wagą neuronu wyjściowego b_m^*, zaś druga – neuronu warstwy ukrytej a_{ij}^* lub σ_{ij}:

a) drugą wagą jest parametr środka funkcji przynależności neuronu warstwy ukrytej a_{ij}^*; do wyznaczenia pochodnej stosujemy wzór (Z2.17),

b) drugą wagą jest parametr szerokości funkcji przynależności neuronu warstwy ukrytej σ_{ij}; do wyznaczenia pochodnej stosujemy wzór (Z2.18).

5. Wyznaczamy drugie pochodne błędu E_k w sytuacji, w której obie wagi, względem których różniczkujemy, są wagami neuronów warstwy ukrytej:

a) jedna z wag jest parametrem środka funkcji przynależności a_{ij}^*, a druga z nich – parametrem szerokości funkcji przynależności σ_{ml}; jeżeli $i \neq m$, czyli są to wagi różnych neuronów, do wyznaczenia pochodnej stosujemy wzór (Z2.27a); jeżeli są to wagi tego samego neuronu, stosujemy wzór (Z2.32),

b) obie wagi są parametrami szerokości funkcji przynależności σ_{ij} i σ_{ml}; jeżeli $i \neq m$, czyli są to wagi różnych neuronów, do wyznaczenia pochodnej stosujemy wzór (Z2.27b); jeżeli są to wagi tego samego neuronu, stosujemy wzór (Z2.34),

c) obie wagi są parametrami środków funkcji przynależności a_{ij}^* i a_{ml}^*; jeżeli $i \neq m$, czyli są to wagi różnych neuronów, do wyznaczenia pochodnej stosujemy wzór (Z2.27c); jeżeli są to wagi tego samego neuronu, stosujemy wzór (Z2.36).

Pamiętajmy również, że macierz hesjanu jest symetryczna, więc w każdym kroku przedstawionego algorytmu musimy wyznaczyć tylko połowę pochodnych. Dla odwrotnego układu wag, względem których różniczkujemy, pochodna będzie miała taką samą wartość jak obliczona.

Z2.3. Wyznaczanie gradientu wyjścia sieci FBF względem wag, dla danego wejścia

W bieżącym punkcie zajmiemy się wyznaczeniem pochodnych wyjścia sieci FBF względem parametrów (wag) modelu, przy danej ustalonej wartości wzorca wejściowego $\mathbf{x} = (x_1, \ldots, x_n)$. Przypomnijmy, że gradient modelu względem współczynników wagowych wykorzystywany jest w rozdziale 3.3.3 do oszacowania wariancji wyjściowej modelu (prognozy otrzymywanej z modelu). Gradienty sieci dla poszczególnych wzorców treningowych stosowane są także przy aproksymacji iloczynem skalarnym (Levenberga–Marquarda) macierzy hesjanu błędu modelu (również patrz rozdział 3.3.3). Oczywiście pochodne sieci względem wag wyliczane są także w sposób niejawny podczas wyznaczania gradientu błędu w algorytmie wstecznej propagacji i w innych metodach, jakie mogą zostać zastosowane do uczenia sieci FBF.

Przyjmujemy, że model prognostyczny dany jest równaniem (Z2.1a) lub (Z2.1b) i, zasadniczo, wszystkie oznaczenia pozostają takie same jak w punkcie Z2.1. Nasze zadanie polega więc na znalezieniu pochodnych wyjścia sieci y względem wag neuronu wyjściowego b_i^*, $i = 1, \ldots, K$ oraz neuronów pierwszej warstwy ukrytej, tj. parametrów środków funkcji przynależności a_{ij}^* i ich szerokości σ_{ij}, $j = 1, \ldots, n$, $i = 1, \ldots, K$.

Ponieważ neuron wyjściowy ma charakter liniowy, więc wyznaczenie pochodnych względem wag tego neuronu b_i^* jest sprawą trywialną. Pochodne te równe są naturalnie wejściom tego neuronu, czyli stanom (wyjściom) odpowiednich neuronów drugiej warstwy ukrytej v_i:

$$\frac{\partial y}{\partial b_i^*} = \frac{\partial}{\partial b_i^*}\left(\sum_{i=1}^{K} v_i b_i^*\right) = v_i \tag{Z2.37}$$

Wyznaczenie pochodnych wyjścia sieci FBF względem wag warstwy ukrytej jest o tyle proste, że zostały one w zasadzie znalezione w punkcie Z2.1. Dokładniej rzecz biorąc, na podstawie wzorów (Z2.8) oraz (Z2.9) pochodną wyjścia sieci FBF względem parametrów środków zbiorów rozmytych poprzedników reguł a_{ij}^*, dla danego ustalonego wzorca treningowego **x**, możemy zapisać:

$$\begin{aligned}\frac{\partial y}{\partial a_{ij}^*} &= \frac{\partial \tau_i}{\partial a_{ij}^*}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) = \\ &= \frac{x_j - a_{ij}^*}{\sigma_{ij}^2}\tau_i\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) = \frac{x_j - a_{ij}^*}{\sigma_{ij}^2}v_i\left(b_i^* - y\right)\end{aligned} \tag{Z2.38}$$

Analogicznie zależności (Z2.8) oraz (Z2.11) pozwalają nam niemal natychmiast otrzymać formułę dla pochodnej wyjścia sieci neuronowo-rozmytej FBF względem parametrów szerokości zbiorów rozmytych poprzedników reguł σ_{ij}, przy danym ustalonym wzorcu treningowym **x**:

$$\begin{aligned}\frac{\partial y}{\partial \sigma_{ij}} &= \frac{\partial \tau_i}{\partial \sigma_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) = \\ &= \frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^3}\tau_i\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(b_i^* - y\right) = \frac{(x_j - a_{ij}^*)^2}{\sigma_{ij}^3}v_i\left(b_i^* - y\right)\end{aligned} \tag{Z2.39}$$

Podsumowując nasze rozważania w bieżącym punkcie, przedstawiony wyżej proces wyznaczania gradientu wyjścia sieci FBF dla danego ustalonego wzorca treningowego **x** przyjmuje postać następującego algorytmu:

1. Podajemy wzorzec $\mathbf{x} = (x_1, \ldots, x_n)$ na wejście sieci FBF. Przy wykorzystaniu zależności (3.7.1a) lub (Z2.1b) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Za pomocą zależności (Z2.37) obliczamy pochodne wyjścia sieci względem wag neuronu wyjściowego b_i^*, $i = 1, \ldots, K$.

3. Za pomocą zależności (Z2.38) obliczamy pochodne wyjścia modelu względem parametrów środków krzywej Gaussa w jego pierwszej warstwie ukrytej, a_{ij}^*, $j = 1,\ldots, n$, $i = 1, \ldots, K$.

4. Za pomocą zależności (Z2.39) obliczamy pochodne wyjścia modelu względem parametrów szerokości krzywej Gaussa w pierwszej warstwie ukrytej sieci, σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$.

Z2.4. Wyznaczanie pochodnych wyjścia sieci FBF względem zmiennych wejściowych

Gradient sieci FBF w przestrzeni wejść wykorzystujemy w punkcie 3.5 do propagacji błędów wejściowych modelu prognostycznego i szacowania związanej z nimi niepewności prognozy. Stosuje się go również w licznych zagadnieniach pozostających poza zakresem tematycznym naszej pracy, takich jak analiza wrażliwości wejść, dobór optymalnej struktury sieci itp. Kolejnym ważnym obszarem zastosowań, w którym wymaga się często wyznaczenia gradientów sieci, są problemy optymalizacji w przestrzeni wejść, w którym predyktor neuronowo-rozmyty stanowi model minimalizowanego (maksymalizowanego) odwzorowania.

Przyjmujemy oczywiście, że model prognostyczny dany jest równaniem (Z2.1a) lub (Z2.1b) i zasadniczo wszystkie oznaczenia pozostają takie same jak w punkcie Z2.1. Naszym zadaniem jest więc wyznaczenie pochodnych wyjścia sieci FBF względem zmiennych wejściowych modelu $\mathbf{x} = (x_1, \ldots, x_n)$.

Gradient wejściowy sieci FBF możemy wyznaczyć w sposób analogiczny do tego w przypadku gradientu błędu, w punkcie Z2.1. Niemal identycznie jak dla wzoru (Z2.5) pochodną wyjścia sieci względem dowolnej zmiennej wejściowej x_j, $j = 1,\ldots, n$ możemy zapisać następująco:

$$\frac{\partial y}{\partial x_j} = \sum_{p=1}^{K} b_p^* \frac{\partial v_p}{\partial x_j} \tag{Z2.40}$$

Dalej, w zbliżony sposób jak w przypadku zależności (Z2.6), ze wzoru na różniczkowanie ilorazu funkcji możemy w następujący sposób wyznaczyć pochodną wyjścia neuronów drugiej warstwy ukrytej v_p względem x_j:

$$\frac{\partial v_p}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{\tau_p}{\sum_{t=1}^{K}\tau_t}\right) = \frac{\frac{\partial \tau_p}{\partial x_j}\sum_{t=1}^{K}\tau_t - \sum_{t=1}^{K}\frac{\partial \tau_t}{\partial x_j}\tau_p}{\left(\sum_{t=1}^{K}\tau_t\right)^2} \tag{Z2.41}$$

Niestety w przeciwieństwie do zależności od parametrów (wag) stan każdego z neuronów pierwszej warstwy ukrytej zależy od wszystkich wejść sieci, więc nie możemy skorzystać z wygodnego uproszczenia wynikającego z zerowania się pochodnych, jakie udało nam się zastosować w przypadku wzoru (Z2.8). Tym niemniej jeśli podstawimy (Z2.41) do (Z2.40), otrzymamy:

$$\begin{aligned}\frac{\partial y}{\partial x_j} &= \sum_{p=1}^{K}\left(b_p^*\frac{\frac{\partial \tau_p}{\partial x_j}\sum_{t=1}^{K}\tau_t - \sum_{t=1}^{K}\frac{\partial \tau_t}{\partial x_j}\tau_p}{\left(\sum_{t=1}^{K}\tau_t\right)^2}\right) = \\ &= \frac{\sum_{t=1}^{K}\tau_t\sum_{p=1}^{K}b_p^*\frac{\partial \tau_p}{\partial x_j} - \sum_{t=1}^{K}\frac{\partial \tau_t}{\partial x_j}\sum_{p=1}^{K}b_p^*\tau_p}{\left(\sum_{t=1}^{K}\tau_t\right)^2} = \\ &= \frac{1}{\sum_{t=1}^{K}\tau_t}\left(\sum_{p=1}^{K}b_p^*\frac{\partial \tau_p}{\partial x_j} - \sum_{t=1}^{K}\frac{\partial \tau_t}{\partial x_j}\sum_{p=1}^{K}b_p^*\frac{\tau_p}{\sum_{t=1}^{K}\tau_t}\right) = \\ &= \frac{1}{\sum_{t=1}^{K}\tau_t}\left(\sum_{p=1}^{K}b_p^*\frac{\partial \tau_p}{\partial x_j} - y\sum_{t=1}^{K}\frac{\partial \tau_t}{\partial x_j}\right) = \frac{1}{\sum_{t=1}^{K}\tau_t}\sum_{p=1}^{K}\frac{\partial \tau_p}{\partial x_j}\left(b_p^* - y\right)\end{aligned} \tag{Z2.42}$$

Pozostaje nam teraz jedynie obliczyć pochodną wyjścia neuronu pierwszej warstwy ukrytej τ_p względem zmiennej wejściowej x_j. Podobnie jak w przypadku wzoru (Z2.9), otrzymujemy następującą zależność:

$$\begin{aligned}\frac{\partial \tau_p}{\partial x_j} &= \frac{\partial}{\partial x_j}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{x_j - a_{pj}^*}{\sigma_{pj}}\right)^2\right) = \\ &= -\frac{x_j - a_{pj}^*}{\sigma_{pj}^2}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{x_j - a_{pj}^*}{\sigma_{pj}}\right)^2\right) = -\frac{x_j - a_{pj}^*}{\sigma_{pj}^2}\tau_p\end{aligned} \tag{Z2.43}$$

Jeżeli teraz podstawimy (Z2.43) do (Z2.42), otrzymamy ostateczną zależność dla pochodnej wyjścia sieci względem zmiennej wejściowej x_j:

$$\frac{\partial y}{\partial x_j} = \frac{1}{\sum_{t=1}^{K} \tau_t} \sum_{p=1}^{K} - \frac{x_j - a^*_{pj}}{\sigma^2_{pj}} \tau_p \left(b^*_p - y \right) = \\ = \sum_{p=1}^{K} \frac{x_j - a^*_{pj}}{\sigma^2_{pj}} \nu_p \left(y - b^*_p \right) \tag{Z2.44}$$

I znów podsumujmy krótko nasze rozważania w bieżącym punkcie, przedstawiając proces wyznaczania gradientu sieci FBF w przestrzeni wejść w postaci następującego algorytmu:

1. Podajemy wzorzec $\mathbf{x} = (x_1, \ldots, x_n)$ na wejście sieci FBF. Przy wykorzystaniu zależności (3.7.1a) lub (Z2.1b) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Pochodne wyjścia sieci względem kolejnych zmiennych wejściowych x_j, $j = 1,\ldots, n$ obliczamy za pomocą zależności (Z2.44).

ZAŁĄCZNIK 3

Ważniejsze gradienty i hesjany związane z siecią neuronowo-rozmytą typu Takagi–Sugeno z liniowymi następnikami reguł

Z3.1. Wyznaczanie gradientu w przestrzeni wag dla błędu sieci neuronowo-rozmytej typu Takagi–Sugeno, przy danym wzorcu treningowym

Sieć neuronowo-rozmytą opartą na modelu wnioskowania typu Takagi–Sugeno z liniowymi funkcjami w następnikach reguł wykorzystywaliśmy do prognoz zapotrzebowania na energię elektryczną w punkcie 2.3.4. Przypomnijmy, że sieć ta stanowi rozszerzenie modelu lingwistycznego FBF, w którym stosuje się iloczynową regułę agregacji wejść oraz wnioskowania wraz z addytywnym scalaniem wyników działania reguł. Funkcje przynależności zbiorów rozmytych A_{ij}, j =1,…, n, i = 1, …, K w wykorzystywanych w poprzednikach poszczególnych reguł systemu mają charakter funkcji Gaussa. Strukturę sieci możemy przedstawić przy użyciu następującego równania:

$$\mathrm{y}(x_1,\ldots,x_n)=\sum_{i=1}^{K}\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-a_{ij}^{*})^2}{\sigma_{ij}^2}\right)(m_{i0}+m_{i1}x_1+\ldots+m_{in}x_n)}{\sum_{i=1}^{K}\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-a_{ij}^{*})^2}{\sigma_{ij}^2}\right)} \tag{Z3.1a}$$

Zauważmy, że równanie modelu (Z3.1a) alternatywnie przedstawić możemy w postaci sieci neuronowej o czterech warstwach węzłów:

$$\tau_i=\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{x_j-a_{ij}^{*}}{\sigma_{ij}}\right)^2\right),\ i=1,\ldots,K,\ \ j=1,\ldots,n$$
$$v_i=\frac{\tau_i}{\sum_{t=1}^{K}\tau_t} \tag{Z3.1b}$$

$$y_i^* = \sum_{j=1}^{K} m_{ij} x_j + m_{i0}$$

$$y = \sum_{i=1}^{K} v_i y_i^*$$

Zauważmy przy tym, że dwie ostatnie warstwy ukryte sieci mają charakter równoległy, wspólnie przekazując sygnał do neuronu wyjściowego. Współczynniki wagowe pierwszej warstwy ukrytej, a_{ij}^* i σ_{ij}, j =1,…, n, i = 1, …, K, są parametrami krzywej Gaussa i mogą być interpretowane jako środki i szerokości funkcji przynależności zbiorów rozmytych poprzedników reguł A_{ij} sieci neuronowo-rozmytej. Natomiast m_{ij}, j = 0,…, n, i = 1, …, K są parametrami funkcji liniowych występujących w następnikach reguł systemu. Druga warstwa ukryta i neuron wyjściowy nie mają bezpośrednich parametrów adaptacyjnych.

W bieżącym punkcie zajmiemy się wyznaczeniem gradientu błędu kwadratowego modelu Takagi–Sugeno względem wag. Jego wyznaczanie ma oczywiście sens jedynie w przypadku sieci trenowanych metodą najmniejszych kwadratów, np. algorytmem wstecznej propagacji błędu przedstawianym w punkcie 2.3.4 (lub innymi metodami gradientowymi minimalizacji błędu kwadratowego modelu). Podobnie jak w przypadku innych rodzajów sieci, pokażemy sposób jego wyznaczania dla danego wzorca treningowego.

Przyjmijmy więc, że zbiór $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, k = 1, …, N stanowi zbiór danych treningowych, dla których wyznaczono błąd E. Zakładamy naturalnie kwadratową postać funkcji błędu daną przez (Z1.3). Dokładnie tak samo jak w przypadku gradientów błędów sieci MLP i FBF (punkt Z1.1 i Z2.1), jeżeli operację różniczkowania wykonujemy dla danego wejściowego wzorca treningowego $\mathbf{x}_k$, to występujące w funkcji błędu stany wyjściowe modelu są stałe, z wyjątkiem jednego zależącego właśnie od $\mathbf{x}_k$, czyli $y(\mathbf{x}_k)$. Ich pochodna jest więc równa zero. W konsekwencji, dla danego $\mathbf{x}_k$ gradient błędu E redukuje się tylko do gradientu składnika błędu E_k odpowiadającego tej obserwacji treningowej:

$$E_k = \frac{1}{2}(y_k - y(x_{k1},\ldots,x_{kn}))^2 = \frac{1}{2}(y_k - y(\mathbf{x}_k))^2 \qquad \text{(Z3.2)}$$

Również i w tym punkcie, w dalszej części naszych rozważań, dla wszystkich pobudzeń i stanów neuronów w sieci wnioskowania Takagi–Sugeno pomijać będziemy dla uproszczenia indeks k, pamiętając jednakże, że wartości te są wyznaczone dla konkretnego wejściowego wzorca treningowego $\mathbf{x}_k$.

Ponieważ model neuronowo-rozmyty (Z3.1) stanowi zasadniczo uogólnienie sieci FBF, w którym wykorzystuje się w następnikach reguł funkcje liniowe wejść zamiast stałych wartości, to większość wyprowadzanych zależności

będzie bardzo zbliżona do wzorów prezentowanych w punkcie Z2.1. Rozpocznijmy od wyznaczenia pochodnych błędu względem wag trzeciej warstwy ukrytej, to jest parametrów funkcji liniowych występujących w następnikach reguł systemu m_{ij}, gdzie $j = 0,\ldots, n$, $i = 1, \ldots, K$. Jako że wyjście sieci zależy od m_{ij} liniowo, bez trudu możemy stwierdzić, że:

$$\frac{\partial E}{\partial m_{ij}} = \frac{\partial E_k}{\partial m_{ij}} = \frac{\partial (\frac{1}{2}(y_k - y)^2)}{\partial m_{ij}} = -(y_k - y)\frac{\partial y}{\partial m_{ij}} = -(y_k - y)\sum_{p=1}^{K} v_p \frac{\partial y_p^*}{\partial m_{ij}} \qquad \text{(Z3.3)}$$

Oczywiście od danej wagi m_{ij} zależny jest tylko stan i-tego neuronu y_i^*, a więc dla pozostałych $p \neq i$ pochodne neuronów y_p^* we wzorze (Z3.3) będą równe zeru. Niemal natychmiast otrzymujemy zatem:

$$\frac{\partial E}{\partial m_{ij}} = -(y_k - y)v_i \frac{\partial y_i^*}{\partial m_{ij}} = -(y_k - y)v_i \frac{\partial}{\partial m_{ij}}\left(m_{i0} + \sum_{t=1}^{n} m_{it} x_{kt} \right) = \qquad \text{(Z3.4)}$$

$$= \begin{cases} -(y_k - y)v_i x_{kj} & \text{dla } j = 1,\ldots, n \\ -(y_k - y)v_i & \text{dla } j = 0 \end{cases}$$

Podobnie znajdziemy zależności dla pochodnych względem parametrów pierwszej warstwy ukrytej a_{ij}^* oraz σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$. Oznaczmy przez w_{ij} dowolną wagę a_{ij}^* lub σ_{ij}. Wówczas oczywiście również dla wag w_{ij} możemy uzależnić pochodną błędu kwadratowego od pochodnej wyjścia sieci dla danego wzorca treningowego $\mathbf{x}_k$:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E_k}{\partial w_{ij}} = -(y_k - y)\frac{\partial y}{\partial w_{ij}} = -(y_k - y)\frac{\partial}{\partial w_{ij}}\left(\sum_{p=1}^{K} v_p y_p^* \right) \qquad \text{(Z3.5)}$$

Zauważmy dalej, że wartości funkcji liniowych następników reguł zależą wyłącznie od parametrów m_{ij} i wejść x_{kj}. Natomiast nie zależą one ani bezpośrednio, ani pośrednio od w_{ij}. W związku z tym stany neuronów trzeciej warstwy ukrytej y_p^* mają względem wag w_{ij} charakter stały. Oznacza to, że pochodną wyjścia sieci neuronowo-rozmytej realizującej wnioskowanie Takagi–Sugeno (Z3.1) względem wagi w_{ij} można znaleźć dokładnie tak samo jak pochodną wyjścia sieci FBF wyznaczoną w punkcie Z2.1. Musimy tylko zastąpić b_p^*, stałe wagi neuronu wyjściowego, również stałymi (względem w_{ij}) stanami neuronów trzeciej warstwy ukrytej y_p^*. Otrzymana dla pochodnej wyjścia sieci FBF zależność (Z2.6) będzie więc obowiązywać również w tym przypadku. Korzystając z niej, mamy:

$$\frac{\partial y}{\partial w_{ij}} = \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \left(y_i^* - y \right) \tag{Z3.6}$$

Neurony pierwszej warstwy ukrytej wykonują dokładnie takie same operacje jak w przypadku sieci FBF, zatem pochodne ich stanów τ_i, względem parametrów a_{ij}^* i σ_{ij}, obliczyć możemy odpowiednio przy użyciu wzorów (Z2.9) i (Z2.11). Ostatecznie więc, podstawiając (Z2.9) do (Z3.6) i dalej korzystając z (Z3.5), formułę dla pochodnej błędu kwadratowego sieci neuronowo-rozmytej (Z3.1) względem parametrów środków zbiorów rozmytych poprzedników a_{ij}^*, przy danej obserwacji treningowej $\mathbf{x}_k$, możemy zapisać w następujący sposób:

$$\frac{\partial E}{\partial a_{ij}^*} = -(y_k - y) \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \nu_i \left(y_i^* - y \right) \tag{Z3.7}$$

I podobnie, również korzystając z wcześniej obliczonych zależności (Z3.5), (Z3.6) i oczywiście w tym przypadku z (Z2.11), otrzymujemy ostateczną formułę dla pochodnej błędu kwadratowego sieci neuronowo-rozmytej Takagi–Sugeno (Z3.1), względem parametrów szerokości zbiorów rozmytych poprzedników reguł systemu σ_{ij}, dla danej obserwacji treningowej $\mathbf{x}_k$:

$$\frac{\partial E}{\partial \sigma_{ij}} = -(y_k - y) \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} \nu_i \left(y_i^* - y \right) \tag{Z3.8}$$

Podsumujmy nasze rozważania w bieżącym punkcie. Przedstawiony w nim proces wyznaczania gradientu błędu sieci neuronowo-rozmytej, realizującej wnioskowanie Takagi–Sugeno (Z3.1), dla danej obserwacji treningowej $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, przyjmuje postać następującego algorytmu:

1. Podajemy wzorzec $\mathbf{x}_k$ na wejście sieci. Wykorzystując równanie modelu dane przez (3.8.1a) lub (Z3.1b), przepuszczamy sygnał przez sieć i obliczamy pobudzenia i stany wszystkich neuronów.

2. Za pomocą zależności (Z3.4) obliczamy pochodne błędu kwadratowego sieci względem parametrów funkcji liniowych występujących w następnikach reguł systemu m_{ij}, gdzie $j = 0,\ldots, n$, $i = 1, \ldots, K$.

3. Za pomocą zależności (Z3.7) obliczamy pochodne błędu kwadratowego względem parametrów środków krzywej Gaussa w pierwszej warstwie ukrytej sieci, a_{ij}^*, $j = 1,\ldots, n$, $i = 1, \ldots, K$.

4. Za pomocą zależności (Z3.8) obliczamy pochodne błędu kwadratowego względem parametrów szerokości krzywej Gaussa w pierwszej warstwie ukrytej sieci, σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$.

Z3.2. Wyznaczanie hesjanu błędu kwadratowego sieci neuronowo-rozmytej typu Takagi–Sugeno względem wag

W bieżącym punkcie zajmiemy się problemem wyznaczenia macierzy **H** drugich pochodnych (hesjanu) funkcji błędu sieci neuronowo-rozmytej systemu rozmytego typu Takagi–Sugeno określonego wzorem (Z3.1) w przestrzeni wag (parametrów modelu). Pamiętajmy, że rozważania tutaj przedstawione mają sens jedynie w przypadku sieci trenowanych metodą najmniejszych kwadratów, np. algorytmem wstecznej propagacji błędu omawianym w punkcie 2.3.4 (lub innymi metodami gradientowymi minimalizacji błędu kwadratowego modelu). Wszystkie oznaczenia stosowane w bieżącym punkcie będą takie same jak w punkcie poprzednim, Z3.1, chyba że zostanie wyraźnie stwierdzone inaczej.

Przypomnijmy, że odwrotność hesjanu błędu wykorzystujemy w przedstawionej w punkcie 3.3.2 metodzie delta do oszacowania macierzy kowariancji parametrów modelu i obliczenia wariancji (lub odchylenia standardowego) warunkowego rozkładu wartości wyjścia modelu dla danego wzorca wejściowego (rozkładu otrzymywanej prognozy). Oczywiście hesjan błędu może być również przydatny do niektórych bardziej zaawansowanych algorytmów uczenia (bądź douczania) sieci neuronowo-rozmytych, w procedurach minimalizacji błędu, w których wykorzystuje się aproksymację kwadratową.

W bieżącym punkcie interesować nas będzie wyłącznie metoda dokładnego wyznaczania hesjanu błędu kwadratowego, zbudowana konkretnie dla struktury sieci neuronowo-rozmytej danej przez (Z3.1). Przypomnijmy tylko, że istnieje szereg metod przybliżonego szacowania macierzy **H**, które mają charakter uniwersalny i mogą być stosowane (przynajmniej do pewnego miejsca) dla modeli różnego typu. W interesującym nas kontekście wyznaczania macierzy kowariancji parametrów modelu najważniejszą z nich jest metoda aproksymacji iloczynem skalarnym (nazywana również aproksymacją Levenberga–Marquarda), którą przedstawiliśmy w punkcie 3.3.2.

Nasz cel polega na znalezieniu drugich pochodnych błędu modelu dla dowolnej pary parametrów. Pamiętajmy jednak, że macierz hesjanu ma charakter symetryczny, więc założenie jakiegoś uporządkowania kolejności parametrów, według których wyznaczane są pochodne, np. związanego z ich położeniem w strukturze sieci neuronowo-rozmytej, nie zmniejsza ogólności rozważań. Dla układu wag w odwrotnej kolejności możemy skorzystać właśnie z symetryczności macierzy **H**.

Przyjmijmy, tak samo jak w poprzednim punkcie, że sieć dopasowana została metodą najmniejszych kwadratów na zbiorze danych $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, $k = 1, \ldots, N$, zakładamy więc, że różniczkowany błąd kwadratowy E (dany tak samo jak w przypadku sieci MLP czy FBF wzorem (Z1.3)) wyznaczono na tym właśnie zbiorze. Całościowy błąd E jest sumą błędów E_k danych przez (Z3.2), związanych z odchyleniami poszczególnych obserwacji treningowych. Zastosujemy więc tę samą procedurę co w punkcie Z1.2 dla sieci MLP czy Z2.2

dla sieci FBF i elementy hesjanu błędu E modelu (Z3.1) będziemy wyznaczać odrębnie, znajdując drugie pochodne poszczególnych składników błędu E_k, a następnie sumując je po wszystkich wzorcach treningowych.

Zajmijmy się obecnie oszacowaniem drugiej pochodnej składnika pojedynczego błędu E_k. Tym razem również, tak jak w poprzednim punkcie, indeks k dla wszystkich pobudzeń i stanów neuronów będziemy dla uproszczenia pomijać, pamiętając jednakże, że ich wartości są wyznaczone dla konkretnego wejściowego wzorca treningowego $\mathbf{x}_k$.

Zdecydowanie najłatwiejsze będzie, rzecz jasna, wyznaczenie drugich pochodnych błędów względem wag trzeciej warstwy ukrytej sieci, czyli współczynników funkcji liniowych występujących w następnikach reguł systemu rozmytego Takagi–Sugeno, to jest m_{ij}, gdzie $j = 0,\ldots, n$, $i = 1, \ldots, K$. Weźmy więc dwa dowolne parametry m_{ij} i m_{dl} i policzmy drugą pochodną błędu E_k względem tych wag. Rozważmy na początek przypadek, w którym $j \neq 0$ (m_{ij} nie jest wyrazem wolnym funkcji linowej). Korzystając z (Z3.4), możemy powyższą pochodną obliczyć następująco:

$$\frac{\partial E_k}{\partial m_{dl}\partial m_{ij}} = \frac{\partial}{\partial m_{dl}}\left(-(y_k - y)v_i x_{kj}\right) = v_i x_{kj}\frac{\partial y}{\partial m_{dl}} =$$

$$= v_i x_{kj} v_d \frac{\partial}{\partial m_{ij}}\left(m_{d0} + \sum_{t=1}^{n} m_{dt}x_{kt}\right) = \begin{cases} v_i v_d x_{kj} x_{kl} & \text{dla } j,l = 1,\ldots,n \\ v_i v_d x_{kj} & \text{dla } l = 0, j = 1,\ldots,n \end{cases} \tag{Z3.9a}$$

Oczywiście dla $j = 0$ otrzymujemy analogicznie:

$$\frac{\partial E_k}{\partial m_{dl}\partial m_{ij}} = \frac{\partial}{\partial m_{dl}}\left(-(y_k - y)v_i\right) = v_i\frac{\partial y}{\partial m_{dl}} =$$

$$= v_i v_d \frac{\partial}{\partial m_{ij}}\left(m_{d0} + \sum_{t=1}^{n} m_{dt}x_{kt}\right) = \begin{cases} v_i v_d x_{kl} & \text{dla } j = 0, l = 1,\ldots,n \\ v_i v_d & \text{dla } j = 0, l = 0 \end{cases} \tag{Z3.9b}$$

I znów, podobnie jak w przypadku wag sieci FBF, jeżeli weźmiemy pod uwagę przedstawiony w punkcie 2.3.4 algorytm znajdowania parametrów m_{ij} metodą regresji liniowej, to wyniki otrzymane w (Z3.9) nie powinny być dla nas specjalną niespodzianką. Wyjście sieci neuronowo-rozmytej Takagi–Sugeno, z liniowymi funkcjami następników reguł, możemy bowiem interpretować jako funkcję liniową transformowanych zmiennych wejściowych $z_{ij} = v_i x_j$ (plus nieco odmiennie traktowany wyraz wolny). W związku z tym bez trudu możemy zauważyć, że (Z3.9) stanowi element obliczeń macierzy kowariancji dla powyższego modelu liniowego, określonej przez zależność (3.2.31).

W kolejnym kroku wyznaczmy pochodne względem parametrów neuronów trzeciej warstwy ukrytej m_{dl}, $l = 0,\ldots, n$, $d = 1, \ldots, K$ oraz pierwszej warstwy ukrytej a_{ij}^* i σ_{ij}, $j = 1,\ldots, n$, $i = 1, \ldots, K$. Oznaczmy, jak w poprzednim punkcie, przez w_{ij} dowolną wagę a_{ij}^* lub σ_{ij}. Ponieważ nie korzystamy tutaj z formuł łańcuchowych do obliczania pochodnych, więc w przeciwieństwie do algorytmu wyznaczania hesjanu błędu sieci perceptronowej MLP (patrz rozdział Z1.2), kolejność wyboru wag podczas różniczkowania funkcji błędu sieci Takagi–Sugeno (Z3.1) nie ma znaczenia. Nieco wygodniej będzie nam zróżniczkować E_k najpierw względem wagi z pierwszej warstwy ukrytej w_{ij}, a następnie m_{dl}. Bez większego trudu można obliczyć, że kolejność odwrotna daje dokładnie ten sam wynik.

Procedura postępowania, którą zastosujemy, będzie bardzo podobna do tej w punkcie Z2.2 dla sieci FBF, co wynika z kilkakrotnie już wskazywanych naturalnych powiązań między rozważanymi modelami neuronowo-rozmytymi. Na podstawie wyznaczonych w poprzednim podrozdziale Z3.1 formuł dla pierwszych pochodnych (wzory (Z3.5) i (Z3.6)) otrzymujemy zatem następującą zależność dla pochodnej błędu E_k względem wagi pierwszej warstwy ukrytej w_{ij}:

$$\frac{\partial E_k}{\partial w_{ij}} = -(y_k - y)\frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}(y_i^* - y) \tag{Z3.10}$$

Jeżeli zależność (Z3.10) zróżniczkujemy następnie względem dowolnej wagi neuronu trzeciej warstwy ukrytej m_{dl}, to otrzymamy formułę dla odpowiedniej drugiej pochodnej błędu kwadratowego E_k. Zauważmy, że obliczenia będą miały zbliżony charakter do wykonanych w punkcie Z2.2 w przypadku wzoru (Z2.16). W zasadzie jedyna różnica polega na zastąpieniu stałej b_i^* wartością funkcji liniowej y_i^* oraz wynikających z tego nieznacznych zmianach w pochodnych funkcji następników sieci. Możemy więc zapisać:

$$\begin{aligned}\frac{\partial^2 E_k}{\partial m_{dl}\partial w_{ij}} &= \frac{\partial}{\partial m_{dl}}\left(-(y_k - y)\frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}(y_i^* - y)\right) = \\ &= \frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\frac{\partial}{\partial m_{dl}}\Big(-(y_k - y)(y_i^* - y)\Big) = \\ &= \frac{\partial \tau_i}{\partial w_{ij}}\frac{1}{\left(\sum_{t=1}^{K}\tau_t\right)}\left(\frac{\partial y}{\partial m_{dl}}(y_i^* - y) - (y_k - y)\left(\frac{\partial y_i^*}{\partial m_{dl}} - \frac{\partial y}{\partial m_{dl}}\right)\right)\end{aligned} \tag{Z3.11}$$

Pochodna wyjścia sieci y względem parametru m_{dl} różni się dla wyrazu wolnego i pozostałych współczynników prostej brakiem lub występowaniem wartości zmiennej wejściowej x_{kl}. Dlatego obecnie również musimy rozważyć obydwa te przypadki oddzielnie. Ponadto zauważmy, że pochodna y_i^* względem m_{dl}, dla $i \neq d$, jest równa zero. Biorąc pod uwagę oba te fakty, możemy (Z3.11) zapisać jako:

$$\frac{\partial^2 E_k}{\partial m_{dl}\partial w_{ij}} = \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(\frac{\partial y}{\partial m_{dl}} (y_i^* - y) - (y_k - y)\left(\frac{\partial y_i^*}{\partial m_{dl}} - \frac{\partial y}{\partial m_{dl}} \right) \right) =$$

$$= \begin{cases} \dfrac{\partial \tau_i}{\partial w_{ij}} \dfrac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} x_{kl} \left(v_d (y_i^* - y) - (y_k - y)(\Delta_{id} - v_d) \right) \text{ dla } l = 1,..., n \\ \dfrac{\partial \tau_i}{\partial w_{ij}} \dfrac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} \left(v_d (y_i^* - y) - (y_k - y)(\Delta_{id} - v_d) \right) \quad \text{dla } l = 0 \end{cases} \tag{Z3.12}$$

gdzie Δ_{id} oznacza deltę Kroneckera, tzn. jest równe 1 dla $i = d$, 0 w przeciwnym przypadku.

Aby obliczyć pochodną wyjścia neuronu pierwszej warstwy ukrytej τ_t względem parametrów a_{ij}^* i σ_{ij}, stosujemy również wzory (Z2.9) i (Z2.11). Ostatecznie więc drugą pochodną E_k względem wag a_{ij}^* i m_{dl} możemy na podstawie (Z2.9) zapisać jako:

$$\frac{\partial^2 E_k}{\partial m_{dl}\partial a_{ij}^*} = \begin{cases} \dfrac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} v_i x_{kl} \left(v_d (y_i^* - y) - (y_k - y)(\Delta_{id} - v_d) \right) \text{ dla } l = 1,..., n \\ \dfrac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} v_i \left(v_d (y_i^* - y) - (y_k - y)(\Delta_{id} - v_d) \right) \quad \text{dla } l = 0 \end{cases} \tag{Z3.13}$$

Podobnie, korzystając z (Z2.11), otrzymujemy zależność dla drugiej pochodnej błędu E_k względem wag σ_{ij} i m_{dl}:

$$\frac{\partial^2 E_k}{\partial m_{dl}\partial \sigma_{ij}} = \begin{cases} \dfrac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} v_i x_{kl} \left(v_d (y_i^* - y) - (y_k - y)(\Delta_{id} - v_d) \right) \\ \qquad\qquad \text{dla } l = 1,..., n \\ \dfrac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} v_i \left(v_d (y_i^* - y) - (y_k - y)(\Delta_{id} - v_d) \right) \text{ dla } l = 0 \end{cases} \tag{Z3.14}$$

Jak zwykle przypadek, gdy obie wagi, względem których wyznaczamy drugą pochodną błędu kwadratowego modelu, pochodzą z pierwszej warstwy ukrytej, jest dużo bardziej złożony. Jeśli jednak przyjrzymy się specyfice sieci neuronowo-rozmytej (Z3.1), implementującej wnioskowanie typu Takagi–Sugeno, to w zasadzie od razu możemy zauważyć, że zależności wyjścia sieci od parametrów poprzedników reguł mają już nie podobny, ale niemal identyczny charakter jak w przypadku sieci FBF w punkcie Z2.2. Porównując równania obu modeli, widzimy, że różnią się one współczynnikami funkcji przetwarzania neuronu wyjściowego. W przypadku sieci FBF były to stałe b_i^*, w sieci Takagi–Sugeno są to wartości funkcji liniowych y_i^*, które również nie zależą od wag pierwszej warstwy ukrytej (a więc są z tego punktu widzenia stałe).

W związku z tym określone w bieżącym punkcie zależności wyznaczania elementów hesjanu błędu dla obu wag z pierwszej warstwy ukrytej będą prawie takie same jak dla sieci FBF – główna różnica polegać będzie na zastąpieniu w otrzymanych równaniach b_i^* przez y_i^*. W bieżącym punkcie ograniczymy się więc tylko do głównego zarysu metody ich znajdowania, odsyłając Czytelnika do szczegółów odpowiednich przekształceń we wzorach z punktu Z2.2.

Weźmy więc dwie dowolne wagi w_{ij} i w_{ml} z pierwszej warstwy ukrytej sieci neuronowo-rozmytej Takagi–Sugeno. Chwilowo nie rozróżniamy, czy są one parametrami środków czy szerokości funkcji Gaussa definiującej funkcje przynależności zbiorów rozmytych poprzedników reguł systemu. Podobnie jak w punkcie Z2.2 dla sieci FBF, jako podstawę algorytmu obliczeń wykorzystamy wzór (3.3.15) uzależniający drugą pochodną błędu kwadratowego sieci od pierwszych i drugich pochodnych jej wyjścia, który przecież zachodzi dla modelu dowolnej postaci.

Możemy więc drugą pochodną błędu E_k zapisać w postaci zależności odpowiadającej formule (Z2.20):

$$\frac{\partial^2 E_k}{\partial w_{ml}\partial w_{ij}} = y'_{ml}y'_{ij} - (y_k - y)y''_{ml,ij} \tag{Z3.15}$$

gdzie $y'_{ml} = \dfrac{\partial y}{\partial w_{ml}}, y'_{ij} = \dfrac{\partial y}{\partial w_{ij}}, y''_{ml,ij} = \dfrac{\partial^2 y}{\partial w_{ml}\partial w_{ij}}$

Analogicznie jak w przypadku wzoru (Z2.21) dla sieci FBF, jedynie korzystając ze wzoru (Z3.6) dla pochodnej wyjścia sieci Takagi–Sugeno, możemy dla pierwszego członu wzoru (Z2.20) zapisać:

$$y'_{ml}\,y'_{ij} = \frac{\partial \tau_m}{\partial w_{ml}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} (y_m^* - y) \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} (y_i^* - y) =$$

$$= \frac{\partial \tau_m}{\partial w_{ml}} \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} (y_m^* - y)(y_i^* - y) \tag{Z3.16}$$

Podobnie jak w przypadku sieci FBF, dla drugiej pochodnej wyjścia sieci $y''_{ml,ij}$, również różniczkując, ale (Z3.6) względem wagi w_{ml}, a następnie wykonując te same przekształcenia, co w (3.7.22) i (Z2.23), otrzymujemy ogólną zależność:

$$y''_{ml,ij} = \frac{\partial}{\partial w_{ml}} \left(\frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} (y_i^* - y) \right) =$$

$$= \frac{\partial^2 \tau_i}{\partial w_{ml} \partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)} (y_i^* - y) - \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \frac{\partial \tau_m}{\partial w_{ml}} \frac{\partial \tau_i}{\partial w_{ij}} (y_i^* + y_m^* - 2y) \tag{Z3.17}$$

Ponieważ wyjście (stan) neuronu warstwy ukrytej τ_i nie zależy od wag innych neuronów, to jego druga pochodna różni się od 0 jedynie, gdy jest liczona względem wag tego samego neuronu. Znów więc, tak samo jak dla sieci FBF, musimy rozważyć odrębnie sytuacje, w której wagi w_{ij} i w_{ml} pochodzą z różnych neuronów oraz pochodzą z tego samego neuronu. W tym pierwszym przypadku, jeżeli $i \neq m$, pierwszy człon (Z3.17) jest równy zero, a więc druga pochodna wyjścia sieci upraszcza się do drugiego członu tego równania:

$$y''_{ml,ij} = -\frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \frac{\partial \tau_m}{\partial w_{ml}} \frac{\partial \tau_i}{\partial w_{ij}} (y_i^* + y_m^* - 2y) \tag{Z3.18}$$

Podstawiając (Z3.18) i (Z3.16) do (Z3.15) oraz wykonując takie same przekształcenia jak w przypadku wzoru (Z2.26) dla modelu FBF, otrzymujemy zależność dla drugiej pochodnej błędu kwadratowego sieci Takagi–Sugeno względem parametrów pochodzących z różnych neuronów pierwszej warstwy ukrytej:

$$\frac{\partial^2 E_k}{\partial w_{ml}\partial w_{ij}} = \frac{\partial \tau_m}{\partial w_{ml}} \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t\right)^2} \left((y_m^* - y)(y_i^* - y) + (y_k - y)(y_i^* + y_m^* - 2y)\right) \quad \text{(Z3.19)}$$

Jak już kilkakrotnie wskazywaliśmy, neurony pierwszej warstwy ukrytej zarówno w przypadku sieci Takagi–Sugeno, jak i sieci FBF, wykonują takie same operacje przetwarzania, a więc pochodne ich wyjść τ_i względem parametrów a_{ij}^* i σ_{ij} obliczone mogą zostać za pomocą tych samych zależności (Z2.9) i (Z2.11). Podstawiając powyższe formuły do (Z3.19), a następnie wykonując te same przekształcenia co w przypadku wzorów (Z2.27), otrzymujemy następujące formuły dla odpowiednich elementów hesjanu błędu E_k względem wag różnych neuronów warstwy ukrytej, tj. dla $i \neq m$:

– różniczkowanie względem a_{ij}^*, a następnie σ_{ml} (przypomnijmy, że w przypadku odwrotnej kolejności wag możemy do obliczenia pochodnej skorzystać z symetryczności macierzy hesjanu):

$$\frac{\partial^2 E_k}{\partial \sigma_{ml}\partial a_{ij}^*} = \frac{(x_{kl} - a_{ml}^*)^2}{\sigma_{ml}^3} \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} v_m v_i \left((y_m^* - y)(y_i^* - y) + (y_k - y)(y_i^* + y_m^* - 2y)\right) \quad \text{(Z3.20a)}$$

– różniczkowanie względem σ_{ij} oraz σ_{ml}:

$$\frac{\partial^2 E_k}{\partial \sigma_{ml}\partial \sigma_{ij}} = \frac{(x_{kl} - a_{ml}^*)^2}{\sigma_{ml}^3} \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} v_m v_i \left((y_m^* - y)(y_i^* - y) + (y_k - y)(y_i^* + y_m^* - 2y)\right) \quad \text{(Z3.20b)}$$

– różniczkowanie względem a_{ij}^* oraz a_{ml}^*:

$$\frac{\partial^2 E_k}{\partial a_{ml}^*\partial a_{ij}^*} = \frac{x_{kl} - a_{ml}^*}{\sigma_{ml}^2} \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} v_m v_i \left((y_m^* - y)(y_i^* - y) + (y_k - y)(y_i^* + y_m^* - 2y)\right) \quad \text{(Z3.20c)}$$

Zajmijmy się teraz przypadkiem, gdy $m = i$, czyli wyznaczamy drugą pochodną błędu kwadratowego E_k modelu względem wag w_{ij} oraz w_{il}, należących

do tego samego neuronu pierwszej warstwy ukrytej. Wówczas, stosując analogiczne przekształcenia jak we wzorze (Z2.28) dotyczącym sieci FBF, zależność (Z3.17) dla drugiej pochodnej wyjścia sieci Takagi–Sugeno względem obu tych wag upraszcza się do postaci:

$$y''_{il,ij} = (y_i^* - y)\left(\frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \frac{\partial^2 \tau_i}{\partial w_{il} \partial w_{ij}} - 2 \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)^2} \frac{\partial \tau_i}{\partial w_{il}} \frac{\partial \tau_i}{\partial w_{ij}} \right) \tag{Z3.21}$$

Podstawiając (Z3.21) i (Z3.16) (z uwzględnieniem faktu że $m = i$) do wzoru (Z3.15), otrzymujemy ogólną formułę, która pozwala na wyznaczenie drugiej pochodnej błędu kwadratowego E_k sieci neuronowo-rozmytej Takagi–Sugeno względem wag w_{ij} oraz w_{il}, należących do tego samego neuronu pierwszej warstwy ukrytej. Również pomijamy tutaj przekształcenia pośrednie, ponieważ będą analogiczne jak w przypadku odpowiedniej zależności dla sieci FBF, określonej wzorem (Z2.30). Jedyna różnica polegać będzie (tak jak w poprzednich zależnościach w bieżącym punkcie) na zastąpieniu b_i^* przez y_i^*.

$$\frac{\partial^2 E_k}{\partial w_{il} \partial w_{ij}} =$$

$$= (y_i^* - y)\left(\frac{\partial \tau_i}{\partial w_{il}} \frac{\partial \tau_i}{\partial w_{ij}} \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)^2} (y_i^* - 3y + 2y_k) - \frac{1}{\left(\sum_{t=1}^{K} \tau_t \right)} \frac{\partial^2 \tau_i}{\partial w_{il} \partial w_{ij}} (y_k - y) \right) \tag{Z3.22}$$

Aby otrzymać konkretne formuły na elementy hesjanu, musimy już tylko w (Z3.22) podstawić właściwe pochodne wyjść neuronów pierwszej warstwy ukrytej τ_i względem odpowiednich wag tych neuronów a_{ij}^* i σ_{ij}. Oczywiście znów możemy tutaj wykorzystać fakt, że neurony pierwszej warstwy ukrytej w przypadku sieci Takagi–Sugeno i sieci FBF wykonują takie same operacje przetwarzania, więc pierwsze pochodne ich wyjść τ_i względem parametrów a_{ij}^* i σ_{ij} obliczone mogą zostać przy użyciu tych samych zależności (Z2.9) i (Z2.11). Podobnie w przypadku drugich pochodnych τ_i, występujących w ostatnim członie (Z3.22), odpowiednie formuły zostały już otrzymane w punkcie Z2.2, jako wzory (Z2.31), (Z2.33) i (Z2.35). Po kolei wyznaczmy więc teraz wzory dla elementów hesjanu błędu kwadratowego E_k, sieci wnioskowana typu Takagi–Sugeno, wynikające z różnych kombinacji a_{ij}^* i σ_{ij}.

Podstawiając więc do (Z3.22) drugą pochodną wyjścia neuronu ze wzoru (Z2.31), zaś pierwsze pochodne z (Z2.9) i (Z2.11), a następnie wykonując analogiczne przekształcenia jak w przypadku wzoru dla sieci FBF (Z2.32), otrzymujemy pochodną względem wag a_{ij}^*, a następnie σ_{il}:

$$\frac{\partial^2 E_k}{\partial \sigma_{il} \partial a_{ij}^*} = \\ = (y_i^* - y) \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{(x_{kl} - a_{il}^*)^2}{\sigma_{il}^3} \nu_i \left(\nu_i (y_i^* - 3y + 2y_k) - (y_k - y) \right) \tag{Z3.23}$$

Dalej, podstawiając drugą pochodną wyjścia neuronu ze wzoru (Z2.33) i pierwsze pochodne z (Z2.9) i (Z2.11) do ogólnego wzoru (Z3.22), a następnie wykonując te same przekształcenia co w przypadku równania (Z2.34) dla sieci FBF, otrzymujemy ostateczną postać drugiej pochodnej względem wag σ_{ij}, a następnie σ_{il}, pochodzących z tego samego neuronu warstwy ukrytej:

$$\frac{\partial^2 E_k}{\partial \sigma_{il} \partial \sigma_{ij}} = \\ = (y_i^* - y) \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ik}^3} \frac{(x_{kl} - a_{il}^*)^2}{\sigma_{il}^3} \nu_i \left(\nu_i (y_i^* - 3y + 2y_k) - (y_k - y) \right) \tag{Z3.24}$$

Pozostała nam do obliczenia już ostatnia pochodna, tym razem względem wag a_{ij}^* i potem a_{il}^*, pochodzących z tego samego neuronu warstwy ukrytej. Otrzymamy ją w analogiczny sposób, podstawiając drugą pochodną wyjścia neuronu ze wzoru (Z2.35) i pierwsze pochodne z (Z2.9) i (Z2.11) do ogólnego wzoru (Z3.22). Ponownie pominiemy przekształcenia otrzymanego wzoru, będą one bowiem niemal identyczne jak w przypadku odpowiedniej zależności dla sieci FBF (Z2.36).

$$\frac{\partial^2 E_k}{\partial a_{il}^* \partial a_{ij}^*} = \\ = (y_i^* - y) \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \frac{x_{kl} - a_{il}^*}{\sigma_{il}^2} \nu_i \left(\nu_i (y_i^* - 3y + 2y_k) - (y_k - y) \right) \tag{Z3.25}$$

Mamy już wobec tego wszystkie elementy potrzebne do wyznaczenia elementów macierzy hesjanu błędu kwadratowego dla modelu sieci neuronowo-rozmytej typy Takagi–Sugeno z liniowymi następnikami reguł, danej przez

(Z3.1). Jak widzimy, algorytm jest dosyć zbliżony do analogicznego rozwiązania z punktu Z2.2, dla sieci FBF. Również i w tym przypadku z powodu różnorodności typów wag występujących w modelu wymaga on zastosowania wielu różnych formuł do obliczenia poszczególnych bloków w macierzy **H**. Przedstawmy go więc w uporządkowanej postaci.

1. Obliczenia wykonujemy oddzielnie dla każdego wzorca danych występującego w zbiorze treningowym $\{\mathbf{x}_k, y_k\} = \{(x_{k1}, \ldots, x_{kn}), y_k\}$, określając wartości elementów hesjanu błędu E_k związanego z tym wzorcem. Odpowiednie pochodne obliczone w kolejnych punktach algorytmu należy więc podsumować dla kolejnych obserwacji treningowych.

2. Podajemy wzorzec $\mathbf{x}_k$ na wejście sieci. Przy wykorzystaniu zależności (3.8.1a) lub (Z3.1b) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

3. Wyznaczamy drugie pochodne błędu E_k w sytuacji, w której obie wagi, względem których różniczkujemy, są wagami trzeciej warstwy ukrytej m_{ij} i m_{dl}. Do obliczenia pochodnej stosujemy wzory (Z3.9a) i (Z3.9b).

4. Wyznaczamy drugie pochodne błędu E_k w sytuacji, w której jedna z wag, względem której różniczkujemy jest wagą trzeciej warstwy ukrytej m_{dl}, zaś druga – neuronu warstwy ukrytej a_{ij}^* lub σ_{ij}:

a) drugą wagą jest parametr środka funkcji przynależności neuronu warstwy ukrytej a_{ij}^*; do wyznaczenia pochodnej stosujemy wzór (Z3.13),

b) drugą wagą jest parametr szerokości funkcji przynależności neuronu warstwy ukrytej σ_{ij}; do wyznaczenia pochodnej stosujemy wzór (Z3.14).

5. Wyznaczamy drugie pochodne błędu E_k w sytuacji, w której obie wagi, względem których różniczkujemy, są wagami neuronów warstwy ukrytej:

a) jedna z wag jest parametrem środka funkcji przynależności a_{ij}^*, a druga z nich – parametrem szerokości funkcji przynależności σ_{ml}; jeżeli $i \neq m$, czyli są to wagi różnych neuronów, do wyznaczenia pochodnej stosujemy wzór (Z3.20a); jeżeli są to wagi tego samego neuronu, stosujemy wzór (Z3.23),

b) obie wagi są parametrami szerokości funkcji przynależności σ_{ij} i σ_{ml}; jeżeli $i \neq m$, czyli są to wagi różnych neuronów, do wyznaczenia pochodnej stosujemy wzór (Z3.20b); jeżeli są to wagi tego samego neuronu, stosujemy wzór (Z3.24),

c) obie wagi są parametrami środków funkcji przynależności a_{ij}^* i a_{ml}^*; jeżeli $i \neq m$, czyli są to wagi różnych neuronów, do wyznaczenia pochodnej stosujemy wzór (Z3.20c); jeżeli są to wagi tego samego neuronu, stosujemy wzór (Z3.25).

Pamiętajmy również, że macierz hesjanu jest symetryczna, a więc w każdym kroku przedstawionego wyżej algorytmu musimy wyznaczyć tylko połowę pochodnych. Dla odwrotnego układu wag, względem których różniczkujemy, pochodna będzie miała taką samą wartość jak obliczona wyżej.

Z3.3. Wyznaczanie gradientu wyjścia sieci neuronowo-rozmytej typu Takagi–Sugeno, dla danego wejścia

Obecnie zajmiemy się wyznaczeniem pochodnych wyjścia sieci neuronowo-rozmytej typu Takagi–Sugeno względem parametrów (wag) modelu, dla ustalonej konkretnej wartości wzorca wejściowego $\mathbf{x} = (x_1, \ldots, x_n)$. Przypomnijmy, że gradient modelu względem współczynników wagowych wykorzystywany jest w rozdziale 3.3.3 do oszacowania wariancji wyjściowej modelu (prognozy otrzymywanej z modelu). Gradienty sieci dla poszczególnych wzorców treningowych stosowane są także przy aproksymacji iloczynem skalarnym (Levenberga–Marquarda) macierzy hesjanu błędu modelu (również patrz rozdział 3.3.3). Pochodne sieci względem wag obliczane są oczywiście w sposób niejawny podczas wyznaczania gradientu błędu w algorytmie wstecznej propagacji i innych metodach, które mogą zostać zastosowane do uczenia sieci.

Przyjmujemy oczywiście, że model prognostyczny dany jest równaniem (Z3.1a) lub (Z3.1b) i zasadniczo wszystkie oznaczenia pozostają takie same jak w punkcie Z2.1. Nasze zadanie polega więc na znalezieniu pochodnych wyjścia sieci y względem wag neuronów trzeciej warstwy ukrytej (parametrów funkcji liniowych występujących w następnikach reguł systemu) m_{ij}, $j = 0,\ldots, n$, $i = 1, \ldots, K$, oraz neuronów pierwszej warstwy ukrytej, tj. parametrów środków funkcji przynależności poprzedników reguł systemu a_{ij}^* i ich szerokości σ_{ij}, $j =1,\ldots, n$, $i = 1, \ldots, K$. Druga warstwa ukryta i neuron wyjściowy nie posiadają bezpośrednich parametrów adaptacyjnych.

Ponieważ odpowiednie pochodne zostały już właściwie niemal w pełni znalezione przy okazji wyznaczania gradientu błędu sieci w przestrzeni wag w punkcie Z3.1, więc tutaj ograniczymy się głównie do podania końcowych wzorów, co do szczegółów odsyłając Czytelnika do odpowiednich miejsc w poprzedzającym tekście.

Rozpocznijmy od wyznaczenia pochodnych wyjścia sieci względem wag trzeciej warstwy ukrytej, to jest parametrów funkcji liniowych występujących w następnikach reguł systemu m_{ij}, gdzie $j = 0,\ldots, n$, $i = 1, \ldots, K$. Pochodne te zostały wyznaczone przy okazji zależności (Z3.3) i (Z3.4). Na ich podstawie otrzymujemy więc:

$$\frac{\partial y}{\partial m_{ij}} = \begin{cases} v_i x_{kj} & \text{dla } j = 1,\ldots,n \\ v_i & \text{dla } j = 0 \end{cases} \tag{Z3.26}$$

Dalej, na podstawie wzoru (Z3.6) oraz zależności na pochodną wyjścia neuronu pierwszej warstwy ukrytej τ_i, względem a_{ij}^* (Z2.9), możemy wyznaczyć pochodną wyjścia sieci neuronowo-rozmytej Takagi–Sugeno względem parame-

trów środków zbiorów rozmytych poprzedników reguł, dla danego ustalonego wzorca treningowego **x**. Niemal natychmiast otrzymujemy następującą formułę:

$$\frac{\partial y}{\partial a_{ij}^*} = \frac{x_{kj} - a_{ij}^*}{\sigma_{ij}^2} \nu_i (y_i^* - y) \tag{Z3.27}$$

Podobnie, również korzystając z zależności (Z3.6) oraz pochodnej τ_i względem σ_{ij} (Z2.11), otrzymujemy z kolei formułę dla pochodnej wyjścia sieci neuronowo-rozmytej Takagi–Sugeno względem parametrów szerokości zbiorów rozmytych poprzedników reguł, przy danym ustalonym wzorcu treningowym **x**:

$$\frac{\partial y}{\partial \sigma_{ij}} = \frac{(x_{kj} - a_{ij}^*)^2}{\sigma_{ij}^3} \nu_i (y_i^* - y) \tag{Z3.28}$$

Podsumowując nasze rozważania w bieżącym punkcie, stwierdzamy, że przedstawiony proces wyznaczania gradientu wejściowego sieci neuronowo-rozmytej wnioskowania typu Takagi–Sugeno (Z3.1), dla danego ustalonego wzorca treningowego **x**, przyjmuje postać następującego algorytmu:

1. Podajemy wzorzec $\mathbf{x} = (x_1, \ldots, x_n)$ na wejście sieci. Przy wykorzystaniu zależności (3.8.1a) lub (Z3.1b), przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Za pomocą zależności (Z3.26) obliczamy pochodne wyjścia sieci względem wag neuronów trzeciej warstwy ukrytej (parametrów funkcji liniowych występujących w następnikach reguł systemu) $m_{ij}, j = 0,\ldots, n, i = 1, \ldots, K$.

3. Za pomocą zależności (Z3.27) obliczamy pochodne wyjścia modelu, względem parametrów środków krzywej Gaussa, w jego pierwszej warstwie ukrytej, a_{ij}^*, $j = 1,\ldots, n, i = 1, \ldots, K$.

4. Za pomocą zależności (Z3.28) obliczamy pochodne wyjścia modelu względem parametrów szerokości krzywej Gaussa, w pierwszej warstwie ukrytej sieci, σ_{ij}, $j = 1,\ldots, n, i = 1, \ldots, K$.

Z3.4. Wyznaczanie pochodnych wyjścia sieci neuronowo-rozmytej typu Takagi–Sugeno względem zmiennych wejściowych

Gradient wejściowy sieci neuronowo-rozmytej implementującej wnioskowanie typu Takagi–Sugeno wykorzystujemy w punkcie 3.5 do propagacji błędów wejściowych modelu prognostycznego i szacowania związanej z nimi niepewności prognozy. Pochodne względem zmiennych wejściowych wykorzystywane są również w wielu zagadnieniach, które pozostają poza zakresem tematycznym naszej pracy, takich jak analiza wrażliwości wejść, dobór opty-

malnej struktury sieci itp. Kolejnym ważnym obszarem zastosowań, w którym wymaga się często wyznaczenia gradientów wejściowych sieci, są problemy optymalizacji w przestrzeni wejść – predyktor neuronowo-rozmyty stanowi tu model minimalizowanego (maksymalizowanego) odwzorowania.

Podobnie jak w poprzednich punktach bieżącego załącznika przyjmujemy, że model prognostyczny dany jest równaniem (Z3.1a) lub (Z3.1b) i wszystkie oznaczenia pozostają takie same jak w punkcie Z3.1. Naszym zadaniem jest więc wyznaczenie pochodnych wyjścia sieci typu Takagi–Sugeno względem zmiennych wejściowych modelu $\mathbf{x} = (x_1, \ldots, x_n)$.

W poprzednich punktach bieżącego załącznika wielokrotnie wykorzystywaliśmy podobieństwa między siecią neuronowo-rozmytą Takagi–Sugeno a siecią FBF. W przypadku pochodnych względem wejść również będziemy mogli zastosować pewne informacje z rozdziału Z2, ale jednak w nieco mniejszym zakresie. W przeciwieństwie do zależności od parametrów, różnice między siecią Takagi–Sugeno i FBF, pod względem zależności wejście–wyjście, są dużo większe.

Zauważmy, że pochodną wyjścia sieci względem dowolnej zmiennej wejściowej x_j, $j = 1,\ldots, n$ możemy zapisać następująco:

$$\frac{\partial y}{\partial x_j} = \sum_{p=1}^{K} y_p^* \frac{\partial v_p}{\partial x_j} + \sum_{p=1}^{K} v_p \frac{\partial y_p^*}{\partial x_j} \tag{Z3.29}$$

W przypadku pierwszego członu (Z3.29) możemy posiłkować się obliczeniami wykonanymi wcześniej dla sieci FBF. Mianowicie jeżeli spojrzymy na zależność (Z2.42), to bez trudu zauważymy, że wykonując te same przekształcenia, zastępując jedynie b_p^* przez y_p^*, otrzymamy:

$$\sum_{p=1}^{K} y_p^* \frac{\partial v_p}{\partial x_j} = \frac{1}{\sum_{t=1}^{K} \tau_t} \sum_{p=1}^{K} \frac{\partial \tau_p}{\partial x_j} (y_p^* - y) \tag{Z3.30}$$

Zwracaliśmy już kilkakrotnie uwagę, że neurony pierwszej warstwy ukrytej sieci neuronowo-rozmytej Takagi–Sugeno wykonują identyczne przetwarzanie jak w przypadku sieci FBF. Oznacza to naturalnie, że pochodną wyjścia τ_p względem zmiennej wejściowej x_j możemy również w obecnym przypadku wyznaczyć według wzoru (Z2.43). Podstawiając (Z2.43) do (Z3.30), otrzymujemy:

$$\sum_{p=1}^{K} y_p^* \frac{\partial v_p}{\partial x_j} = \frac{1}{\sum_{t=1}^{K} \tau_t} \sum_{p=1}^{K} -\frac{x_j - a_{pj}^*}{\sigma_{pj}^2} \tau_p (y_p^* - y) = \\ = \sum_{p=1}^{K} \frac{x_j - a_{pj}^*}{\sigma_{pj}^2} v_p (y - y_p^*) \tag{Z3.31}$$

Wyznaczenie pochodnej w drugim członie (Z3.29) stanowi zadanie elementarne, jako że funkcje następników reguł y_p^* mają charakter liniowy. Możemy zatem napisać:

$$\sum_{p=1}^{K} v_p \frac{\partial y_p^*}{\partial x_j} = \sum_{p=1}^{K} v_p \frac{\partial}{\partial x_j} \left(\sum_{l=1}^{n} m_{pl} x_l \right) = \sum_{p=1}^{K} v_p m_{pj} \tag{Z3.32}$$

Ostatecznie więc na podstawie (Z3.31) i (Z3.32) otrzymujemy zależność dla pochodnej wyjścia sieci względem zmiennej wejściowej x_j:

$$\frac{\partial y}{\partial x_j} = \sum_{p=1}^{K} \frac{x_j - a_{pj}^*}{\sigma_{pj}^2} v_p (y - y_p^*) + \sum_{p=1}^{K} v_p m_{pj} = \\ = \sum_{p=1}^{K} v_p \left(\frac{x_j - a_{pj}^*}{\sigma_{pj}^2} (y - y_p^*) + m_{pj} \right) \tag{Z3.33}$$

Na koniec podsumujmy krótko nasze rozważania w bieżącym punkcie, przedstawiając proces wyznaczania gradientu sieci neuronowo-rozmytej typu Takagi–Sugeno w przestrzeni wejść w postaci następującego algorytmu:

1. Podajemy wzorzec $\mathbf{x} = (x_1, \ldots, x_n)$ na wejście sieci FBF. Przy wykorzystaniu zależności (Z3.1a) lub (Z3.1b) przepuszczamy sygnał przez sieć, obliczając pobudzenia i stany wszystkich neuronów.

2. Pochodne wyjścia sieci względem kolejnych zmiennych wejściowych $x_j, j = 1,\ldots, n$ obliczamy za pomocą zależności (Z3.33).

Literatura

Bakirtzis A.G., Theocharis J.B., Kiartzis S.J., Satsios K.J. (1995), *Short term load forecasting using fuzzy neural networks*, „IEEE Transactions on Power Systems", vol. 10, no. 3, s. 1518–1524.

Bardzki W., Bartkiewicz W. (1995), *Artificial neural networks for load forecasting in transformation period*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk, J. Gorzkowski, A. Nieczaj (eds), *Present-Day Problems of Power Engineering (APE '95)*, Gdańsk–Jurata, s. 15–22.

Bardzki W., Bartkiewicz W., Zieliński J.S. (1995), *Zastosowanie metod sztucznej inteligencji w elektroenergetyce*, [w:] J. Nazarko, W. Zalewski (red.), *Materiały I Ogólnopolskiego Sympozjum Naukowego „Systemy ekspertowe, sieci neuronowe i zbiory rozmyte w elektroenergetyce"*, Białystok, s. 19–26.

Bardzki W., Bartkiewicz W., Gontar Z., Zieliński J.S. (1998), *Short-term electrical load forecasting in transformation period (case study)*, [w:] M. Heiss (ed.), *International ICSC/IFAC Symposium on Neural Computation – NC '98*, Vienna, s. 324–328.

Bardzki W., Bartkiewicz W., Gontar Z., Zieliński J.S. (1999), A *survey of short-term load forecasting algorithms for transient period*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk (eds), *Present-Day Problems of Power Engineering (APE '99)*, vol. 5, Gdańsk–Jurata, s. 11–18.

Bartkiewicz W. (1996), *Fault tolerance of the neural predictor in real world forecasting problems*, [w:] *Neural Networks and Their Applications*, Szczyrk, s. 19–24.

Bartkiewicz W. (1998a), *Krótkoterminowe prognozowanie obciążenia sieci elektroenergetycznej z wykorzystaniem podejścia neuronowo-rozmytego*, [w:] *Sieci i systemy informatyczne – teoria, projekty, wdrożenia*, Łódź, s. 157–166.

Bartkiewicz W. (1998b), *Metody sztucznej inteligencji w prognozowaniu obciążenia sieci elektroenergetycznej*, rozprawa doktorska, Uniwersytet Łódzki, Łódź (maszynopis).

Bartkiewicz W. (1998c), *Short-term load forecasting using fuzzy neural network*, [w:] *Colloquia in Artificial Intelligence (CAI '98)*, Łódź, s. 155–163.

Bartkiewicz W. (1999a), *Analysis of the neural predictor for short-term load forecasting problems*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk (eds), *Present Day Problems of Power Engineering (APE '99)*, vol. 5, Gdańsk–Jurata, s. 19–26.

Bartkiewicz W. (1999b), *Confidence intervals for neural predictor for short-term load forecasting problems*, [w:] *Neural Networks and Their Applications*, Zakopane, s. 649–655.

Bartkiewicz W. (1999c), *Sieci neuronowe w krótkoterminowym prognozowaniu obciążeń sieci elektroenergetycznej dla potrzeb wspomagania decyzji*, [w:] *Systemy Wspomagania Organizacji* (SWO '99), Ustroń, s. 335–344.

Bartkiewicz W. (2000a), *Confidence intervals prediction for the short-term electrical load neural forecasting models*, „Elektrotechnik und Informationstechnik", no. 1(117), s. 8–12.

Bartkiewicz W. (2000b), *Distribution of the neural predictor residuals for short-term load forecasting problems*, [w:] L. Rutkowski, R. Tadeusiewicz (eds), *Neural Networks and Soft Computing*, Zakopane, s. 112–117.

Bartkiewicz W. (2000c), *Impact of the temperatures prediction uncertainty on the short-term load forecasting accuracy*, [w:] *Colloquia in Artificial Intelligence (CAI '2000)*, Łódź, s. 29–36.

Bartkiewicz W. (2000d), *Neuro-fuzzy approach to short-term electrical load forecasting*, [w:] S.-I. Amari, C.L. Giles, M. Gori, V. Piuri (eds), *Neural Computing: New Challenges and Perspectives for the New Millenium*, vol. 6, Proceedings of the International Joint Conference on Neural Networks, IJCNN2000, Como, Italy, s. 229–234.

Bartkiewicz W. (2000e), *Wpływ niepewności wejść na dokładność predyktora neuronowego w zagadnieniach krótkoterminowej prognozy obciążeń sieci elektroenergetycznej*, [w:] *Sieci i systemy informatyczne – teoria, projekty, wdrożenia, aplikacje*, Łódź, s. 187–196.

Bartkiewicz W. (2001a), *Error bars for short term load forecasting neural networks models*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk (eds), *Present-Day Problems of Power Engineering (APE '2001)*, vol. 3, Gdańsk–Jurata, s. 75–82.

Bartkiewicz W. (2001b), *Impact of the input uncertainty on short-term load forecasting accuracy for neural networks models*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk (eds), *Present-Day Problems of Power Engineering (APE '2001)*, vol. 3, Gdańsk–Jurata, s. 67–74.

Bartkiewicz W. (2002), *Neural network stock price predictors and trading decisions risk*, [w:] *Artificial Intelligence in Control and Management*, Łódź, s. 111–118.

Bartkiewicz W. (2011a), *Metody określania niepewności prognoz krótkoterminowego obciążenia sieci dla modeli neuronowych i neuronowo-rozmytych*, „Rynek Energii", nr 1(92), s. 41–46.

Bartkiewicz W. (2011b), *Metody wyznaczania przedziałów prognozy dla rodziny neuronowo rozmytych modeli krótkoterminowego prognozowania obciążenia sieci*, [w:] XI Międzynarodowa Konferencja Naukowa „Prognozowanie w elektroenergetyce", Wisła, 14–16 września, materiały niepublikowane.

Bartkiewicz W. (2011c), *Short-term load forecasting with neuro-fuzzy models*, [w:] *Present-Day Problems of Power Engineering (APE '2011)*, vol. 3, Gdańsk–Jurata, s. 65–72.

Bartkiewicz W. (2012), *Prediction intervals for short-term load forecasting neuro-fuzzy models*, „Przegląd Elektrotechniczny", nr 10b, s. 284–287.

Bartkiewicz W., Bolek C., Gontar B., Gontar Z., Krygier N., Matusiak B., Pamuła A., Papińska-Kacperek J., Zieliński J.S. (2010), *Zastosowania sztucznej inteligencji w elektroenergetyce w pracach Katedry Informatyki UŁ*, [w:] J. Gołuchowski, B. Filipczyk (red.), *Wiedza i komunikacja w innowacyjnych organizacjach. Systemy ekspertowe – wczoraj, dziś, jutro*, Wydawnictwo AE w Katowicach, Katowice, s. 250–256.

Bartkiewicz W., Butkevych O.F., Kyrylenko O.V., Levitskiy V.G., Pavlovskiy V.V., Zieliński J.S. (2001), *Hybrid systems in electric power systems*, [w:] *Zastosowania komputerów w elektrotechnice*, Poznań–Kiekrz, s. 203–206.

Bartkiewicz W., Czajkowska R., Głuszkowski T., Gontar B., Gontar Z., Krygier N., Kurzyjamski R., Matusiak B., Pamuła A., Papińska-Kacperek J., Zieliński J.S. (2004), *Modern IT tools in management*, „Acta Universitatis Lodziensis. Folia Oeconomica", z. 178, s. 7–28.

Bartkiewicz W., Gontar Z., Matusiak B., Pamuła A., Zieliński J.S. (2004), *Control and management in energy market upon deregulation*, „Tekhnichna Elektrodynamika", no. 1, s. 128–133.

Bartkiewicz W., Gontar Z., Matusiak B., Zieliński J.S. (2001a), *Neural network based short-term load forecasting for energy market*, [w:] J.S. Zieliński, K. Ciach (eds), *Energy Market*, Katedra Informatyki Uniwersytetu Łódzkiego, Łódź, s. 73–83.

Bartkiewicz W., Gontar Z., Matusiak B., Zieliński J.S. (2001b), *Zastosowanie narzędzi sztucznej inteligencji w zarządzaniu w elektroenergetyce*, [w:] W. Błaszczyk, B. Kaczmarek (red.), *Przeszłość i przyszłość nauk o zarządzaniu. Zarządzanie, modele, koncepcje i strategie*, Katedra Zarządzania Uniwersytetu Łódzkiego, Łódź, s. 363–374.

Bartkiewicz W., Gontar Z., Matusiak B., Zieliński J.S. (2002), *Short-term load forecasting in market environment*, [w:] *Proceedings of III-d IEE Mediterranean Conference and Exhibi-*

tion on Power Generation, Transmission, Distribution and Energy Conversion (Med Power), Athens, materiały na CD.

Bartkiewicz W., Gontar Z., Matusiak B., Zieliński J.S., Chmielewski M., Szady S. (2002), *Experiences from initial exploitation of the short-term energy demand forecasting system in Zamość Energy Corporation S.A.*, [w:] *Modern Electric Power Systems*, Wrocław, s. 59–63.

Bartkiewicz W., Gontar Z., Zieliński J.S., Bardzki W. (2000a), *Neural-heuristic approach to short-term electrical load forecasting problems*, [w:] H. Bothe, R. Rojas (eds), *Neural Computation (NC'2000)*, Berlin, s. 740–744.

Bartkiewicz W., Gontar Z., Zieliński J.S., Bardzki W. (2000b), *Uncertainty of the short-term load forecasting in utilities*, [w:] S.-I. Amari, C.L. Giles, M. Gori, V. Piuri (eds), *Neural Computing: New Challenges and Perspectives for the New Millenium*, Proceedings of the International Joint Conference on Neural Networks, IJCNN2000, vol. 6, Como, Italy, s. 235–240.

Bartkiewicz W., Matusiak B. (2003), *Short-term load forecasting for energy markets*, [w:] L. Rutkowski, J. Kacprzyk (eds), *Neural Networks and Soft Computing*, Berlin–Heidelberg, s. 790–795.

Bartkiewicz W., Matusiak B. (2004), *Sieci neuronowe i algorytmy genetyczne a krótkookresowe prognozowanie zużycia na rynku energii*, [w:] J.S. Zieliński (red.), *Jerzy S. Zieliński – 50 lat pracy naukowej*, Łódź, s. 345–353.

Bartkiewicz W., Zieliński J.S. (1998), *Zastosowania narzędzi sztucznej inteligencji do prognozowania zapotrzebowania na energię elektryczną*, [w:] *Komputerowo zintegrowane zarządzanie*, Zakopane, s. 27–34.

Belina Z., Węgliński J., Zieliński J.S. (1996), *Sterowanie popytem na energię*, „Biuletyn Informacyjny PTPiREE", nr 4.

Bishop C.M. (1992), *Exact calculation of the Hessian matrix for the multi-layer perceptron*, „Neural Computation", no. 4, s. 494–501.

Bishop C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford.

Brandt J. (2012), *Projekt PCR. Market Coupling*, materiały konferencyjne Forum Obrotu, Giżycko, http://www.polpx.pl/fm/upload/Prezentacje-Forum-Obrotu-2012/4_Projekt_PCR_ForumObrotu_ JB_cze2012.pdf (dostęp: 25.05.2013).

Brandt S. (1998), *Analiza danych. Metody statystyczne i obliczeniowe*, Warszawa.

Buntine W.L., Weigand A.S. (1994), *Computing second derivatives in feed-forward networks: A review*, „IEEE Transaction on Neural Networks", vol. 5, no. 3.

Butkevych O.F., Pawłowskiy W.W., Bartkiewicz W., Zieliński J.S. (2002), *Hybrid systems in power systems solving*, „Tekhnichna Elektrodynamika", no. 3, s. 77–82.

Chen S.-T., Yu D.C., Moghaddamjo A.R. (1992), *Weather sensitive short-term load forecasting using nonfully connected artificial neural network*, „IEEE Transactions on Power Systems", vol. 7, no. 3, s. 1098–1105.

Chryssolouris G., Lee M., Ramsey A. (1996), *Confidence interval prediction for neural network models*, „IEEE Transactions on Neural Networks", vol. 7, no. 1, s. 229–232.

Cook E. (1971), *The Flow of Energy in an Industrial Society*, „Scientific American", no. 225(3), s. 135–142.

Dash P.K., Dash S., Rahman S. (1993), *A fuzzy adaptive correction scheme for short term load forecasting using fuzzy layered neural network*, [w:] Y. Tamura, H. Suzuki, H. Mori (eds), *Proceedings of the Second International Forum on Applications of Neural Networks to Power Systems*, Yokohama, s. 432–437.

da Silva A.P.A., Moulin L.S. (2000), *Confidence intervals for neural network based short-term load forecasting*, „IEEE Transactions on Power Systems", vol. 15, no. 4, s. 1191–1196.

David H.A., Nagaraja H.N. (2003), *Order Statistics*, Hoboken, New Jersey.

Dietl M., Makowski K. (2010), *Monopolizacja, demonopolizacja, niepewność*, „Biblioteka Regulatora", Urząd Regulacji Energetyki, Warszawa, http://www.ure.gov.pl/download.php?s=1&id=3023 (dostęp: 25.05.2013).

Dillon T.S., Sestito S., Leung S. (1991), *An adaptive neural network approach in load forecasting in a power system*, [w:] *Proceedings of the First International Forum on Applications of Neural Networks to Power Systems*, Seattle, s. 17–21.

Ding A.A. (1999), *Neural-network prediction with noisy predictors*, „IEEE Transactions on Neural Networks", vol. 10, no. 5, s. 1196–1203.

Draper N.R., Smith H. (1973), *Analiza regresji stosowana*, Warszawa.

Drezga I., Rahman S. (1999), *Short-term load forecasting with local ANN predictors*, „IEEE Transactions on Power Systems", vol. 14, no. 3, s. 844–850.

Durlik I. (1992), *Organizacja i zarządzanie produkcją*, Warszawa.

Durlik I. (1995), *Inżynieria zarządzania. Strategia i projektowanie systemów produkcyjnych*, cz. I, Warszawa.

Durlik I. (1996), *Inżynieria zarządzania. Strategia i projektowanie systemów produkcyjnych*, cz. II, Warszawa.

Dybowski R., Roberts S.J. (1999), *Confidence Intervals and Prediction Intervals for Feed-forward Neural Networks. Technical Report*, Kings College, London, http://www.robots.ox.ac.uk/~sjrob/Pubs/rdsrnnerr.ps.gz (dostęp: 25.05.2013).

Efron B., Tibshirani R. (1993), *An Introduction to the Bootstrap*, New York.

El-Sharkawi M.A., Oh S., Marks R.J. II, Damborg M.J., Brace C.M. (1991), *Short term electric load forecasting using an adaptively trained layered perceptron*, [w:] *Proceedings of the First International Forum on Applications of Neural Networks to Power Systems*, Seattle, s. 3–6.

Erkmen I., Ozsokmen H.V. (1993), *A hybrid neural network fix short-term load forecasting*, [w:] *Proceedings Joint International Power Conference (Power Tech)*, vol. 2, Athens, s. 811–815.

Gontar Z., Hatziargyriou N. (2001), *Short term load forecasting with radial basis function network*, [w:] J.T. Saraiva, M.A. Matos (eds), *IEEE Porto Power Tech Proceedings*, Porto.

Gontar Z., Sideratos G., Hatziargyriou N. (2004), *Short-term load forecasting using radial basis function networks*, [w:] *Third Hellenic Conference on AI Methods and Applications of Artificial Intelligence*, s. 432–438.

[GPW-REK 2010] Giełda Papierów Wartościowych w Warszawie SA, *Szczegółowe zasady obrotu na rynku dobowo-godzinowym energii elektrycznej (REK GPW)*, tekst ujednolicony według stanu prawnego na dzień 3 grudnia 2010 r., http://www.poee.gpw.pl/download/rek (dostęp: 25.05.2013).

Hanmandlu M., Chauhan B.K. (2011), *Load forecasting using hybrid models*, „IEEE Transactions on Power Systems", vol. 26, no. 4, s. 20–29.

Heine S., Malko J., Mikołajczak H., Skorupski W. (1994), *Sieci neuronowe w zagadnieniach systemu elektroenergetycznego na przykładzie modelowania procesu zapotrzebowania mocy*, materiały XII Krajowej Konferencji Automatyki, Gdynia.

Hertz J., Krogh A., Palmer R.G. (1993), *Wstęp do teorii obliczeń neuronowych*, Warszawa.

Heskes T. (1997), *Practical confidence and prediction intervals*, [w:] M. Mozer, M. Jordan, T. Petsche (eds), *Advances in Neural Information Processing Systems 9 (NIPS 97)*, Cambridge, s. 176–182.

Ho K.-L., Hsu Y.-Y., Yang C.-C. (1992), *Short term load forecasting using a multilayer neural network with an adaptive learning algorithm*, „IEEE Transactions on Power Systems", vol. 7, no. 1, s. 141–149.

Hornik K., Stinchcombe M., White H. (1989), *Multilayer feedforward networks are universal approximators*, „Neural Networks", vol. 2, s. 359–366.

Hsu Y.-Y., Yang C.-C. (1991a), *Design of artificial neural networks for short-term load forecasting. I. Self-organising feature maps for day type identification. Generation, transmission and distribution*, „IEE Proceedings C", vol. 138, no. 5, s. 407–413.

Hsu Y.-Y., Yang C.-C. (1991b), *Design of artificial neural networks for short-term load forecasting. II. Multilayer feedforward networks for peak load and valley load forecasting. Generation, transmission and distribution*, „IEE Proceedings C", vol. 138, no. 5, s. 414–418.

Huang C.M., Yang H.-T. (2001), *Evolving wavelet-based networks for short-term load forecasting*, „IEE Proceedings-Generation, Transmission and Distribution", vol. 148 , no. 3, s. 222–228.

Huntsberger T.L., Ajjimarangsee P. (1990), *Parallel self-organising feature maps for unsupervised pattern recognition*, „International Journal of General Systems", vol. 16, no. 4, s. 357–372.

Jabłoński W.J., Bartkiewicz W. (2006), *Systemy informatyczne zarządzania. Klasyfikacja i charakterystyka systemów*, Bydgoszcz.

Jang J.-S.R., Sun C.-T. (1993), *Functional equivalence between Radial Basis Function Networks and Fuzzy Inference Systems*, „IEEE Transactions on Neural Networks", vol. 4, no. 1, s. 156–159.

Jang J.-S.R., Sun C.-T., Mizutani E. (1997), *Neurofuzzy and soft computing*, Upper Saddle River.

Khosravi A., Nahavandi S., Creighton D. (2010), *Construction of optimal prediction intervals for load forecasting problems*, „IEEE Transactions on Power Systems", vol. 25, no. 3, s. 1496–1503.

Kim K.-H., Park K.-J., Hwang J.-K., Kim S.-H. (1995), *Implementation of hybrid short-term load forecasting system using artificial neural networks and fuzzy expert systems*, „IEEE Transactions on Power Systems", vol. 10, no. 3, s. 1534–1539.

Korbicz J., Obuchowicz A., Uciński D. (1994), *Sztuczne sieci neuronowe – podstawy i zastosowania*, Warszawa.

Lambert-Torres G., Traore C.O., Mandolesi F.G., Mukhedkar D. (1991), *Short-term load forecasting using a fuzzy engineering tool*, [w:] *Proceedings of the First International Forum on Applications of Neural Networks to Power Systems*, Seattle, s. 36–40.

Lee K.Y., Cha Y.T., Ku C.C. (1991), *A study on neural networks for short-term load forecasting*, [w:] *Proceedings of the First International Forum on Applications of Neural Networks to Power Systems*, Seattle, s. 26–30.

Lichota A. (2006), *Prognozowanie krótkoterminowe na lokalnym rynku energii elektrycznej*, rozprawa doktorska, Akademia Górniczo-Hutnicza im. Stanisława Staszica, Kraków (maszynopis).

Lin C.-T., Lee C.S.G. (1996), *Neural Fuzzy Systems: A Neuro-fuzzy Synergism to Intelligent Systems*, Upper Saddle River, NJ.

MacKay D.J.C. (1991), *A practical Bayesian framework for backprop networks*, „Neural Computation", vol. 4, no. 3, s. 415–447.

MacKay D.J.C. (1994), *Probable networks and plausible predictions – A review of practical Bayesian methods for supervised neural networks*, http://www.inference.phy.cam.ac.uk/mackay/network.pdf (dostęp: 25.05.2013).

MacKay D.J.C. (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge.

Malko J. (1995), *Wybrane zagadnienia prognozowania w elektroenergetyce*, Wrocław.

Malko J., Mikołajczak H., Skorupski W. (1995), *Artificial neural network based models for short- and long-term load forecasting in the power system*, [w:] *Proceedings of the IEEE/KTH Stockholm Power Tech Conference*, Stockholm.

Marshall K.T., Oliver R.M. (1995), *Decision Making and Forecasting*, New York.

Masters T. (1995), *Neural, Novel and Hybrid Algorithms for Time Series Prediction*, New York.

Masters T. (1996), *Sieci neuronowe w praktyce. Programowanie w języku C++*, Warszawa.

Mastorocostas P.A., Theocharis, J.B., Bakirtzis A.G. (1999), *Fuzzy modeling for short term load forecasting using the orthogonal least squares method*, „IEEE Transactions on Power Systems", vol. 14, no. 1, s. 390–396.

Matusiak B., Bartkiewicz W. (2001), *Linking neural predictors with decision models: The short--term load forecasting case study*, [w:] Z. Szczerba, L. Olbrych, R. Pochyluk (eds), *Present--Day Problems of Power Engineering (APE '2001)*, vol. 3, Gdańsk–Jurata, s. 109–116.

Michalski D., Krysta B., Lelątko P. (2004), *Zarządzanie ryzykiem na rynku energii*, Warszawa.

Midera A. (2011), *Aktywny odbiorca energii elektrycznej na rynku bilansującym w Polsce*, „Elektroenergetyka – Współczesność i Rozwój", nr 4(10), s. 10–16.

Mielczarski W. (2000), *Rynki energii elektrycznej. Wybrane aspekty techniczne i ekonomiczne*, Warszawa.

Mielczarski W. (ed.) (2005), *Development of Electricity Markets*, Łódź.

Moody J.E. (1992), *The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems*, [w:] J.E. Moody, S.J. Hanson, R.P. Lippmann (eds), *Advances in Neural Information Processing Systems 4 (NIPS 91)*, San Mateo, s. 847–854.

Mori H., Kobayashi H. (1995), *Optimal fuzzy inference for short-term load forecasting*, [w:] *Proceedings of the IEEE Power Industry Computer Application Conference*, Salt Lake City, s. 312–318.

Muhlemann A.P., Oakland J.S., Lockyer K.G. (1995), *Zarządzanie, produkcja i usługi*, Warszawa.

Nazarko J., Jurczuk A. (1996), *Zarządzanie zapotrzebowaniem na energię elektryczną na poziomie rejonu i zakładu energetycznego*, [w:] Z. Połecki (red.), *Rynek energii elektrycznej: Rynek hurtowy, rynki lokalne. Materiały III Konferencji Naukowo-Technicznej*, Nałęczów.

Neal R.M. (1996), *Bayesian Learning for Neural Networks*, New York.

Okólski M. (red.) (2001), *Jaki model rynku energii?*, „Biblioteka Regulatora", Warszawa, http://www.ure.gov.pl/portal/pl/217/ Jaki_model_rynku_energii.html (dostęp: 25.05.2013).

Pandey A.S., Singh D., Sinha S.K. (2010), *Intelligent hybrid wavelet models for short-term load forecasting*, „IEEE Transactions on Power Systems", vol. 25, no. 3, s. 1266–1273.

Park D.C., El-Sharkawi M.A., Marks R.J., Atlas L.E., Damborg M.J. (1991), *Electric load forecasting using an artificial neural network*, „IEEE Transactions on Power Systems", vol. 6, no. 2.

Paska J. (2010), *Elektroenergetyka w Polsce – Od monopolu do konsolidacji?*, „Rynek Energii", nr 4(89), s. 9–17.

[PE-JT-URE 2011] *Prawo energetyczne. Ustawa z dnia 10 kwietnia 1997 r.*, tekst ujednolicony w Biurze Prawnym Urzędu Regulacji Energetyki na dzień 10 kwietnia 2011.

[PE-JT-URE 2012] *Prawo energetyczne. Ustawa z dnia 10 kwietnia 1997 r.* (Dz.U. z 2012, poz. 1059 j.t.), tekst ujednolicony w Biurze Prawnym Urzędu Regulacji Energetyki na dzień 25 września 2012.

Peng T.M., Hubele N.F., Karady G.G. (1990), *Conceptual approach to the application of neural network for short-term load forecasting*, [w:] *IEEE International Symposium on Circuits and Systems*, vol. 4, New Orleans, s. 2942– 2945.

Peng T.M., Hubele N.F., Karady G.G. (1992), *Advancement in the application of neural networks for short-term load forecasting*, „IEEE Transactions on Power Systems", vol. 7, no. 1, s. 250–257.

Penny W.D., Roberts S.J. (1997), *Neural network predictions with error bars*, „Research Report" TR-97-1, Dept. of Electrical and Electronic Engineering, Technology and Medicine, Imperial College of Science, London, http://www.robots.ox.ac.uk/~sjrob/Pubs/nnerrors.ps.gz (dostęp: 25.05.2013).

Petiau B. (2009), *Confidence interval estimation for short-term load forecasting*, [w:] *Proceedings of the PowerTech IEEE Conference*, Bucharest, s. 1–6.

Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. (1992), *Numerical Recipes in C*, Cambridge.

[PSE-IRiESP 2012] Polskie Sieci Elektroenergetyczne Operator SA, *Instrukcja ruchu i eksploatacji sieci przesyłowej. Bilansowanie systemu i zarządzanie ograniczeniami systemowymi*, tekst jednolity obowiązujący od dnia 1 lutego 2013 r., http://www.pse-operator.pl/uploads/kontener/IRiESP-Bilansowanie_tekst_jednolity_01022013_po_KA_CB_7_2012.pdf (dostęp: 25.05.2013).

[PSE-WIRE-STD 2010] Polskie Sieci Elektroenergetyczne Operator SA, *Standardy techniczne systemu WIRE*, wersja 11.0 (aktualizacja), Warszawa, 9 lipca 2010, http://www.pse-operator.pl/uploads/kontener/ Standardy_techniczne_systemu_WIRE_wer_11_0_aktualizacja.pdf (dostęp: 25.05.2013).

Rahman S., Drezga I., Rajagopalan J. (1993), *Knowledge enhanced connectionist models for short-term electric load forecasting*, [w:] Y. Tamura, H. Suzuki, H. Mori (eds), *Proceedings of the Second International Forum on Applications of Neural Networks to Power Systems*, Yokohama, s. 401–406.

Rahman S., Hazim O. (1993), *A generalized knowledge-based short-term load-forecasting technique*, „IEEE Transactions on Power Systems", vol. 8, no. 2, s. 508–514.

Ranaweera D.K., Hubele N.F., Papalexopoulos A.D. (1995), *Application of radial basis function neural network model for short-term load forecasting*, „IEE Proceedings – Generation, Transmission and Distribution", vol. 142, no. 1, s. 45–50.

Refenes A.-P.N. (ed.) (1995), *Neural Networks in the Capital Markets*, Chichester.

[RO-PLAT 2010] *Rozporządzenie Ministra Gospodarki z dnia 17 września 2010 r. w sprawie określenia sposobu i trybu organizowania i przeprowadzania przetargu na sprzedaż energii elektrycznej oraz sposobu i trybu sprzedaży energii elektrycznej na internetowej platformie handlowej* (Dz.U., nr 186, poz. 1246).

Rutkowska D. (1997), *Inteligentne systemy obliczeniowe. Algorytmy genetyczne i sieci neuronowe w systemach rozmytych*, Warszawa.

Rutkowska D., Piliński M., Rutkowski L. (1997), *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte*, Warszawa.

Song H.S., Wang X.F. (2003), *Operation of Market-oriented Power Systems*, London.

Srinivasan D., Liew A.C., Chen J.S.P. (1991), *Short term forecasting using neural network approach*, [w:] *Proceedings of the First International Forum on Applications of Neural Networks to Power Systems*, Seattle, s. 12–16.

Stair R.M. (1992), *Principles of Information Systems – A Managerial Approach*, Thomson Publishing, Boston.

Szczygieł L. (2001), *Model rynku energii elektrycznej*, [w:] M. Okólski (red.), *Jaki model rynku energii?*, „Biblioteka Regulatora", Urząd Regulacji Energetyki, Warszawa, http://www.ure.gov.pl/portal/pl/217/ Jaki_model_rynku_energii.html (dostęp: 25.05.2013).

Taylor J.W. (2012), Short-term load forecasting with exponentially weighted methods, „IEEE Transactions on Power Systems", vol. 27, no. 1, s. 458–464.

[TGE 2012] Towarowa Giełda Energii SA, *Regulamin obrotu rynku towarów giełdowych Towarowej Giełdy Energii SA*, z dnia 13 września 2012 r., http://www.polpx.pl/fm/upload/27092012__Regulamin_RTG.pdf (dostęp: 25.05.2013).

[TGE-RDB 2010] Towarowa Giełda Energii SA, *Szczegółowe zasady obrotu i rozliczeń dla energii elektrycznej na rynku dnia bieżącego*, z dnia 8 grudnia 2010 r., weszły w życie z dniem 15 grudnia 2010 r., http://www.polpx.pl/fm/upload/Rynek-RDB/Szczegowe_zasady_obrotu_i_rozlicze_RDB.pdf (dostęp: 25.05.2013).

[TGE-RDN 2012] Towarowa Giełda Energii SA, *Szczegółowe zasady obrotu i rozliczeń dla energii elektrycznej na rynku dnia następnego*, z dnia 29 maja 2012 r., weszły w życie z dniem 11 czerwca 2012 r., http://www.polpx.pl/fm/upload/RDN/11062012Szczegowe_zasady_obrotu_i_rozlicze_RDN.pdf (dostęp: 25.05.2013).

[TGE-RPM 2012] Towarowa Giełda Energii SA, *Rynek praw majątkowych*, materiały opublikowane w witrynie internetowej Towarowej Giełdy Energii SA, http://www.tge.pl/fm/upload/Wszystko-o-RPM/ FolderRPM.pdf (dostęp: 19.09.2012).

Tibshirani R. (1996), *A comparison of some error estimates for neural network models*, „Neural Computation", no. 8, s. 152–163.

Tong H. (1990), *Non-linear Time Series*, Oxford.

Tresp V., Hofman R. (1998), *Nonlinear time-series prediction with missing and noisy data*, „Neural Computation", vol. 10, s. 731–747.

Tresp V., Neuneier R., Ahmad S. (1995), *Efficient methods for dealing with missing data in supervised learning*, [w:] G. Tesauro, D.S. Touretzky, T.K. Leen (eds), *Advances in Neural Information Processing Systems 7*, Cambridge, MA.

[UOKiK 2011] Urząd Ochrony Konkurencji i Konsumentów, *Pozycja konsumenta na rynku energii elektrycznej. Raport UOKiK*, Warszawa–Wrocław, http://www.uokik.gov.pl/download.php?plik=10178 (dostęp: 25.05.2013).

[URE 2011] Urząd Regulacji Energetyki, *Raport Krajowy Prezesa Urzędu Regulacji Energetyki 2011*, Warszawa, http://www.ure.gov.pl/download/1/4527/Raport_2011_na_strone.pdf (dostęp: 25.05.2013).

[URE 2012] Urząd Regulacji Energetyki, *Sprawozdanie z działalności Prezesa URE w 2011 roku*, „Biuletyn Urzędu Regulacji Energetyki", nr 2(80).

Wang L.-X., Mendel J.M. (1992), *Fuzzy basis functions, universal approximation and orthogonal least squares learning*, „IEEE Transactions on Neural Networks", vol. 3, no. 5, s. 807–815.

Wang Y., Xia Q., Kang C. (2011), *Secondary forecasting based on deviation analysis for short-term load forecasting*, „IEEE Transactions on Power Systems", vol. 26, no. 2, s. 500–507.

White H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge.

Witkowski T. (2011), *Energia – możliwości naukowe i bariery technologiczne oraz społeczne*, „Czysta Energia", nr 5, 2011.

Wright W.A. (1999), *Neural Network Regression with Input Uncertainty*, NCRG/99/008 (raport techniczny), Neural Computing Research Group, Aston University.

Wu H.-C., Lu C.-N. (1999), *Automatic fuzzy model identification for short-term load forecast*, „IEE Proceedings – Generation, Transmission, Distribution", vol. 146, no. 5, s. 477–482.

Yager R.R., Filev D.P. (1995), *Podstawy modelowania i sterowania rozmytego*, Warszawa.

Ying H. (1998a), *General SISO Takagi–Sugeno fuzzy systems with linear rule consequent are universal approximators*, „IEEE Transactions on Fuzzy Systems", vol. 6, no. 4, s. 582–587.

Ying H. (1998b), *General Takagi–Sugeno fuzzy systems are universal approximators*, [w:] *Proceedings of the 1998 IEEE World Congress on Computational Intelligence*, the 1998 IEEE International Conference on Fuzzy Systems, vol. 1, Anchorage, s. 819–823.

Zadeh L.A. (1965), *Fuzzy sets*, „Information Control", vol. 8, s. 338–353.

Zapranis A., Refenes A.-P. (1999), *Principles of Neural Model Identification, Selection and Adequacy*, London.

Zeng X.J., Singh M.G. (1995), *Approximation theory of fuzzy systems – MIMO case*, „IEEE Transactions on Fuzzy Systems", vol. 3, no. 2, s. 219–235.

Zerka M. (2001), *Zarządzanie ryzykiem na konkurencyjnym rynku energii elektrycznej*, Materiały CIRE, http://www.cire.pl/publikacje/Art_Zerka.pdf (dostęp: 25.05.2013).

Zerka M. (2003), *Strategie na rynkach energii elektrycznej*, Warszawa.

Zhang Y., Zhou Q., Sun C., Lei S., Liu Y., Song Y. (2008), *RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment*, „IEEE Transactions on Power Systems", vol. 23, no. 3, s. 853–858.

Zieliński J.S. (red.) (2000), *Inteligentne systemy w zarządzaniu – Teoria i praktyka*, Warszawa.

Żurada J.M., Barski M., Jędruch W. (1996), *Sztuczne sieci neuronowe*, Warszawa.

Spis rysunków i tabel

Od Redakcji

Witold Bartkiewicz jest wieloletnim pracownikiem Katedry Informatyki Uniwersytetu Łódzkiego na Wydziale Zarządzania UŁ, z którą związał się w 1987 r. Od 2005 r. pełni w niej obowiązki kierownika Zakładu Sztucznej Inteligencji i Narzędzi Informatyki. Jego wykształcenie i doświadczenie zawodowe łączy elementy informatyki oraz zarządzania. Studia ukończył na Uniwersytecie Łódzkim, otrzymując dyplom magistra matematyki o specjalności metody numeryczne i programowanie. Stopień doktora uzyskał także na Uniwersytecie Łódzkim – w zakresie nauk o zarządzaniu. Tytuł rozprawy doktorskiej brzmiał: *Metody sztucznej inteligencji w prognozowaniu obciążenia sieci elektroenergetycznej*.

W okresie tym główny obszar zainteresowań badawczych doktora W. Bartkiewicza stanowiła problematyka zastosowań narzędzi informatycznych, a przede wszystkim sztucznej inteligencji – sieci neuronowych, systemów z logiką rozmytą oraz systemów ekspertowych – w zarządzaniu elektroenergetyką. Jest on autorem i współautorem kilkudziesięciu publikacji poświęconych tej dziedzinie. W większości prac koncentruje się przy tym na zagadnieniach inteligentnych metod krótkoterminowego prognozowania zapotrzebowania na energię elektryczną oraz wspomagania problemów decyzyjnych występujących w przedsiębiorstwach elektroenergetycznych i zasilanych przez powyższe prognozy.

Doktor W. Bartkiewicz brał również udział w wielu międzynarodowych i krajowych projektach dotyczących wykorzystania narzędzi informatycznych w zarządzaniu przedsiębiorstwami elektroenergetycznymi. Wymienić można tutaj: projekt celowy nr 45/CS6-9/93 „Opracowanie projektu, wykonanie i uruchomienie oprogramowania systemu informatycznego dla kierownictwa w zakładzie energetycznym", projekt SJEP-07149-94 Escadina „Nowe techniki w sterowaniu elektroenergetyką" w ramach programu Tempus, projekt celowy nr 8T 10B 035 97C/3430 „Analiza narzędzi sztucznej inteligencji i opracowanie metody prognozowania krótkoterminowego zapotrzebowania na energię elektryczną przy użyciu wybranych narzędzi komputerowych", program SYNERGY PLUS, temat „Expanding the Competition Intelligence in the European Distributed Energy Resources Sector" w ramach PR6 (numer kontraktu ETI-CT-2005-

-023395). Jest również współautorem kilku wykorzystywanych w praktyce systemów informatycznych związanych ze wspomaganiem zarządzania elektroenergetyką.

Witold Bartkiewicz ma także wieloletnie doświadczenia dydaktyczne w nauczaniu zastosowań narzędzi informatycznych i sztucznej inteligencji w biznesie. Brał udział w opracowaniu kilku podręczników akademickich dotyczących systemów informatycznych zarządzania, baz danych, społeczeństwa informacyjnego oraz zastosowań sztucznej inteligencji w zarządzaniu.