UNIWERSYTET EKONOMICZNY W POZNANIU

Wydział Informatyki
i Gospodarki Elektronicznej

Natalia Wojciechowska

Segmentacja gospodarstw domowych ze względu na zużycie energii elektrycznej

Segmentation of households based on the electricity consuption

Praca licencjacka

Promotor: dr Maciej Beręsewicz

Pracę przyjęto dnia:

Podpis promotora

Kierunek: Informatyka i ekonometria

Specjalność: Analityka gospodarcza

Poznań 2019

Spis treści

Wstęp

Szybki rozwój gospodarki oraz wzrost poziomu i jakości życia społeczeństwa wymaga zapewnienia coraz większych dostaw energii elektrycznej. Strategicznym wyzwaniem dla gospodarek poszczególnych państw jest zapewnienie trwałych dostaw energii dla wszystkich odbiorców, modernizacja sposobów jej pozyskiwania i dostarczania do klientów oraz troska o środowisko naturalne. Z uwagi na bezwzględną potrzebę wykorzystania energii elektrycznej w procesach produkcyjnych i konsumpcji bardzo trudno wyobrazić sobie jak świat mógłby wyglądać i funkcjonować bez prądu, który obecnie wykorzystywany jest niemalże w każdym aspekcie życia.

Ogromnym znaczeniem i wyzwaniem jest więc dbanie o bezpieczeństwo rynku energii elektrycznej. W celu jego zapewnienia podejmowanych i wprowadzanych jest wiele planów, narzędzi oraz technik. Dotyczą one zarówno mikro jak i makroregionów, są wprowadzane w obszarze dzielnic, miast, krajów a nawet całego świata. Szanse na usprawnienie i zagwarantowanie bezpieczeństwa na rynku energii elektrycznej daje wprowadzenie, doskonalenie i przede wszystkim nieustanne rozpowszechnianie inteligentnych sieci elektroenergetycznych. Z ich rozwojem związany jest również znaczny wzrost danych generowanych w sektorze rynku energii elektrycznej. Daje to szanse na przeprowadzanie różnego rodzaju badań i analiz, które wcześniej, bez dokładnych pomiarów nie były możliwe.

Tematem pracy jest zastosowanie analizy skupień do grupowania gospodarstw domowych ze względu na ilość zużywanej energii elektrycznej. Celem niniejszej pracy jest próba wyodrębnienia jednorodnych podzbiorów z obiektów badanej populacji oraz zbadanie w jaki sposób nowe źródło danych może zostać wykorzystane do celów przeprowadzenia oficjalnych statystyk.

Badanie zostało wykonane na danych informujących o ilości zużytej energii elektrycznej w kWh, przez 5 490 gospodarstw domowych z Londynu w lutym 2013r. Na potrzebę badania wybrano i wykorzystano rzeczywiste dane pochodzące z portalu *The London Datastore*, który jest prowdzony przez władze miasta Londyn. Odczyty ilości zużywanych kWh w każdym gospo-

darstwie były wykonywane w półgodzinnych interwałach, co dało łącznie 48 zapisów z każdej doby. Próba biorąca udział w badaniu została sklasyfikowana jako zrównoważona reprezentacja populacji mieszkańców Londynu.

Poniża praca składa się z trzech rozdziałów. Pierwszy z nich został poświęcony rynkowi energii elektrycznej w Unii Europejskiej i Wielkiej Brytanii oraz rozwojowi inteligentnych sieci elektroenergetycznych i czujników zbierających informacje o zużyciu prądu.

W drugim rozdziale zostały przedstawione teoretyczne podstawy analizy skupień oraz szczegółowo opracowane metody wykorzystane do późniejszych obliczeń.

Trzeci rozdział zawiera opis zmiennych wykorzystanych do badania, przeprowadzone kolejno obliczenia oraz interpretacje wyników i prezentacje wniosków wynikających z przeprowadzonej analizy.

Teoretyczna część pracy powstała w oparciu o literaturę z zakresu rynku energii elektrycznej, statystyki oraz analizy i wizualizacji danych. Wszelkie obliczenia zostały przeprowadzone w programie R w oparciu o autorski kod przygotowany specjalnie do poniższego badania.

Rozdział 1

Rynek energii elektrycznej

1.1 Funkcjonowanie rynku energii elektrycznej i jego istota

Sektor energetyczny odgrywa bardzo dużą rolę we wzroście i rozwoju gospodarczym każdego państwa. Potocznie używane i zrozumiałe dla większości osób pojęcie rynku energii nie ma pełnej definicji w terminologii ekonomicznej. Swoje wyjaśnienie zawdzięcza założeniu, że energię można traktować jak każde inne dobro oraz, że wszystkie występujące prawa ekonomii znajdują w jej sektorze swoje zastosowanie. Zgodnie z tym, gospodarkę energii elektrycznej traktujemy jako rynek, który określa ogół zawieranych transakcji kupna i sprzedaży dóbr i usług oraz całego szeregu towarzyszących im determinantów (Tarnawski & Młynarski, 2016).

W przedstawionej definicji możemy wyróżnić dwa zasadnicze elementy: (1) popyt reprezentowany przez nabywcę oraz (2) podaż reprezentowaną przez zbywcę, a także niewymienioną cenę, czyli wartość przy której zarówno kupujący jak i sprzedający są gotowi dokonać danej wymiany. Ta właśnie kwota staje się kołem napędowym mechanizmu rynkowego, ponieważ wyższe ceny powodują zmniejszenie ilości nabywanych towarów oraz zachęcają do zwiększenia produkcji, z kolei niższe ceny pobudzają wśród klientów konsumpcjonizm i skłaniają do zmniejszenia ilości wytwarzanych dóbr (Niedziółka, 2010).

Mierzony ceną i wielkością energii pozostającej do dyspozycji, rynek energii elektrycznej jest obiektem zainteresowania zarówno odbiorców indywidualnych jak i wielkich przedsiębiorstw. W obecnych czasach ma on ogromne znaczenie, czego dowodem jest fakt, że korporacje działające na tym rynku są zaliczane do największych podmiotów gospodarczych na świecie (Niedziółka, 2018).

Wyróżniamy dwa podstawowe typy energii: (1) energię pierwotną oraz (2) energię wtórną.

Na energię pierwotną składają się m.in. ropa naftowa, węgiel, gaz ziemny, energia słoneczna oraz energia spadku wód czy pływów morskich, czyli źródła energii, które człowiek wykorzystuje kolejno w procesach przetwórstwa przemysłowego, w celu otrzymania energii elektrycznej. Występowanie tych źródeł na świecie jest bardzo zróżnicowane, w skutek czego pozwala państwom na rozwój gospodarczy, postęp techniczny i wzrost konkurencyjności. Z kolei energię wtórną tworzy energia przetworzona, uzyskana dzięki wykorzystaniu w procesie technologicznym wspomnianych wcześniej źródeł energii. Na całym świecie jest ona niezbędnym warunkiem prowadzenia życia codziennego i działalności gospodarczej. Ilość i rodzaj produkowanej energii jest miernikiem poziomu rozwoju ekonomicznego, zagospodarowania infrastrukturalnego oraz potencjalnych możliwości dalszego rozwoju danego kraju (Michalski, 2005).

Będąca głównym tematem pracy energia, zarówno pierwotna jak i wtórna, stanowiąca towar, którego pozyskanie lub wytwarzanie jest celem działania ogromnych korporacji transnarodowych, stała się impulsem do powstania rynku energii elektrycznej. Historia tworzenia i kształtowania się tego rynku nie jest długa, jest powiązana z narodzeniem się popytu na energię elektryczną oraz zapotrzebowaniem na ciepło pochodzące ze źródeł scentralizowanych i paliwo do silników spalinowych. Systemy elektroenergetyczne w ich obecnym kształcie rozwinęły się dopiero po drugiej wojnie światowej, początkowo rynek tworzyły silnie scentralizowane struktury pozostające w rękach państwa, przybierające postać monopolu (Bielecki, 2007).

Wynikało to głównie z dwóch przesłanek. Po pierwsze, wytwarzanie energii i jej przemysł wymagał ogromnych nakładów finansowych i innowacyjności, a państwo mogło ówcześnie być ich gwarantem. Po drugie, charakter dostaw przesądzał o połączeniu monopolistycznej metody organizacji zaopatrzenia w dobra i usługi infrastrukturalne z publicznym nadzorem. Wynikało to również ze skłonności państw Europy Zachodniej do centralizacji produkcji, a państw Europy Środkowo-Wschodniej do tendencji koncentracji całej gospodarki w rękach państwa. Wówczas na rynku energii elektrycznej jeden przedsiębiorca był odpowiedzialny za produkcję energii, a drugi za jej przemysł. Taki układ pozwalał na osiągnięcie korzyści skali i w sposób naturalny eliminował potrzebę powstawania konkurencji. Oprócz tych zalet, wytworzyły się również zjawiska negatywne, m.in. wadliwy mechanizm racjonalizujący alokację kapitału i pracy oraz arbitralne stanowienie cen. Wynikiem takich działań był wzrost kosztów produkcji energii, co bezpośrednio wywołało znaczące zwiększenie kosztów dla konsumentów. W konsekwencji tych wydarzeń zrodziła się potrzeba znalezienia innych warunków funkcjonowania rynku energii. Szczególnie było to zauważalne w krajach wysokorozwiniętych, które jako pierwsze odczuwają

nowe trendy i wyznaczają kierunki zmian (Niedziółka, 2018).

Do głównych celów przeprowadzanych reform zalicza się chęć: polepszenia efektywności ekonomicznej dostaw energii elektrycznej, zagwarantowania bezpieczeństwa energetycznego, zmniejszenia cen energii elektrycznej dla odbiorców finalnych, a także wprowadzenie innowacyjnych technologii, które poprawiłyby jakość świadczonych usług i w efekcie uczuliły rynek na konkurencję. (Sobierajski & Wilkosz, 2000)

Jednym z najważniejszych kroków czynionych w kierunku rozwoju rynku energii była liberalizacja elektroenergetyki. Pod jej pojęciem rozumie się wszystkie reformy dotyczące funkcjonowania tej gałęzi gospodarki (Niedziółka, 2018):

- komercjalizację – przeobrażenie w firmy zorientowane rynkowo państwowych przedsiębiorstw
- prywatyzację – sprzedaż prywatnym firmom aktywów przedsiębiorstw państwowych
- deregulację – zmniejszenie lustracji administracyjnej i bezpośredniej kontroli nad działalnością przedsiębiorstw elektroenergetycznych
- wprowadzenie konkurencji – pozwolenie na niezależny wybór producenta energii elektrycznej oraz uprawnienie konsumentów do wyboru ze zwiększonej liczby potencjalnych dystrybutorów dostawcy energii elektrycznej i usług elektroenergetycznych

Wymienione reformy występowały współbieżnie z różnym nasileniem lub niezależnie od siebie. W każdym państwie miały inną strukturę ze względu na: format i uwarunkowania krajowej elektroenergetyki, warunki ekonomiczne, architekturę własności, bogactwo w surowce naturalne i strukturę ich zużycia do produkcji energii elektrycznej. Oprócz tego ogromne znaczenie w realizacji reform miały też inne czynniki, m.in. postęp technologiczny i rewolucja teleinformatyczna, nowoczesne technologie komputerowe, technika światłowodowa, ogniwa fotowoltaniczne i turbiny zasilane energią wiatru. Ponadto na zachodzące na rynku zmiany miała wpływ również rozwijająca się konkurencja w skali globalnej, oddziałująca na globalizację rynków energii elektrycznej. Przedsiębiorstwa po uzyskaniu dostępu do rynków zagranicznych mogły zacząć w nie inwestować oraz pozyskiwać krajowe firmy. W efekcie tego utworzyły się ogólnoświatowe standardy i struktury organizacyjne rynku energii elektrycznej (Niedziółka, 2018).

Z uwagi na bezwzględną potrzebę wykorzystania energii elektrycznej w procesach produkcyjnych i konsumpcji, modernizacja jej fizycznej dostawy i kosztów wytworzenia stała się strategicznym wyzwaniem dla gospodarek poszczególnych państw. Z tego powodu, głównym celem

funkcjonowania rynku energii elektrycznej jest zapewnienie adekwatnych cen zarówno dla odbiorców indywidualnych jak i odbiorców biznesowych. Za pośrednictwem wdrożenia mechanizmów rynkowych zmierza się do zmniejszenia kosztów produkcji oraz dostaw energii elektrycznej, co w sposób ostateczny ma wpłynąć na spadek ceny za energię. W związku z tym bardzo ważne jest stworzenie odpowiedniej do warunków, jakie posiada dany kraj, struktury rynku energii elektrycznej. W państwach wysoko rozwiniętych, które mają zdolności podażowe i przemysłowe większe od zapotrzebowania na energię zwykle odczuwa się szybki spadek cen. Natomiast w państwach rozwijających się, systemy elektroenergetyczne często nie nadążają za rozwojem gospodarki, w skutek czego ceny energii mogą wzrastać. Należy wówczas podjąć nowoczesne inwestycje w elektroenergetyce, które zwiększą moce produkcyjne, by ceny tego dobra zaczęły spadać (Pach-Gurgul, 2012).

Podsumowując, celami funkcjonowania rynku energii elektrycznej są przede wszystkim (Pach-Gurgul, 2012):

- zapewnienie bezpieczeństwa dystrybucji energii elektrycznej,
- polepszenie stopnia zasilania w energię elektryczną,
- dążenie do unowocześnienia elektroenergetyki, zainwestowania w nowoczesne źródła energii oraz wytwarzania energii w sposób przyjazny dla środowiska naturalnego,
- zagwarantowanie wszystkim konsumentom lepszej ochrony, poprzez wprowadzenie wielu reform prawnych oraz regulacji rynkowych,
- umożliwienie klientom dowolnego wyboru dostawcy energii elektrycznej,
- zapewnienie rentowności przedsiębiorstwom działającym na rynku energii elektrycznej,
- zobowiązanie do zapewnienia energetyce środków pozwalających na rozwój infrastruktury technicznej.

1.2 Charakterystyka rynku energii elektrycznej w Unii Europejskiej

Polityka energetyczna Unii Europejskiej przez lata podlegała głębokim przemianom. Na jej cele i środki wpływały różnorodne czynniki, najważniejsze z nich to: światowe kryzysy ener-

getyczne, sytuacja gospodarczo–polityczna krajów członkowskich oraz zagrożenia ekologiczne. W wyniku tych ewolucji ukształtowały się jej trzy nadrzędne cele: konkurencyjność gospodarki, bezpieczeństwo energetyczne państw należących do Unii Europejskiej a także ochrona środowiska naturalnego przed negatywnymi skutkami wytwarzania i dystrybucji energii. Powstanie w 1951r. Europejskiej Wspólnoty Węgla i Stali (EWWiS) jest uważane za symboliczny początek polityki energetycznej Unii Europejskiej (Wojtkowska-Łodej, 2014).

Począwszy od lat dziewięćdziesiątych XX wieku, nastąpił wzrost działań w kierunku wspólnej polityki energetycznej, ze szczególnym uwzględnieniem norm prawnych oraz poprawy przejrzystości cen gazu i energii elektrycznej. Unia Europejska poszerzała swoją działalność głównie w obszarze energetyki, ochrony ludności i turystyki. Przyczyniała się do budowy i rozwoju sieci transeuropejskich, infrastruktury transportowej, telekomunikacji i energetyki (Paska & Surma, 2013).

Reagując na gwałtowny wzrost cen ropy naftowej, który miał miejsce na początku 1999r. Unia Europejska opracowała dokument, zwierający m.in. słabości unijnej energetyki i założenia mające zagwarantować bezpieczeństwo energetyczne krajów członkowskich. Wskazywał on gównie na zwiększające się uzależnienie Unii Europejskiej od ropy naftowej i niezadowalające wyniki dotychczasowej polityki kontroli konsumpcji energii w krajach Wspólnoty. Sformułowano ważne założenia długofalowej strategii energetycznej, w których: uwzględniono możliwość podjęcia kontroli zapotrzebowania energetycznego w państwach, podkreślono znaczenie instrumentów podatkowych w kontroli zużycia energii oraz walkę z globalnym ociepleniem uznano za priorytet (Michalski, 2004).

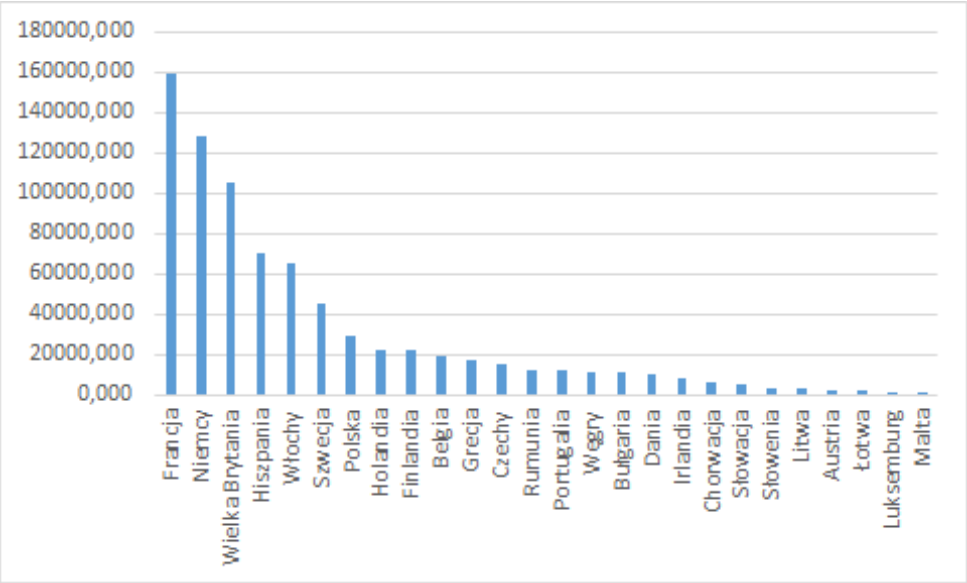
Współczesny kształt polityki energetycznej ściśle powiązany jest z ustaleniami dotyczącymi otwierania krajowych rynków na konkurencję, powiększania międzynarodowego, wewnątrzunijnego handlu energią oraz zapewnienia Europie bezpieczeństwa dostaw surowców energetycznych, zwłaszcza w sytuacjach awaryjnych. Od roku 2007 państwa członkowskie i konsumenci przemysłowi mogą dobrowolnie wybierać swojego dostawcę energii elektrycznej i gazu spośród szerokiego grona konkurentów (Swora & Muras, 2010).

Ostatnia próba zjednolicenia rynku energii całej Wspólnoty została podjęta przez Parlament Europejski w 2009r., kiedy to przyjęto trzeci pakiet liberalizacyjny regulujący (Pach-Gurgul, 2012):

- rozdzielenie działalności obrotowej i wytwórczej od przemysłowej w poszczególnych państwach

- zwiększenie praw konsumentom i ochronę najbardziej wrażliwych odbiorców
- rozpowszechnienie innowacyjnych systemów pomiarowych
- umocnienie uprawnień regulacyjnych
- określenie warunków dotyczących wejścia na europejski rynek energii firm spoza Unii Europejskiej

Zużycie energii elektrycznej w wybranych krajach Unii Europejskiej w roku 2016 przestawiono na wykresie 1.1.

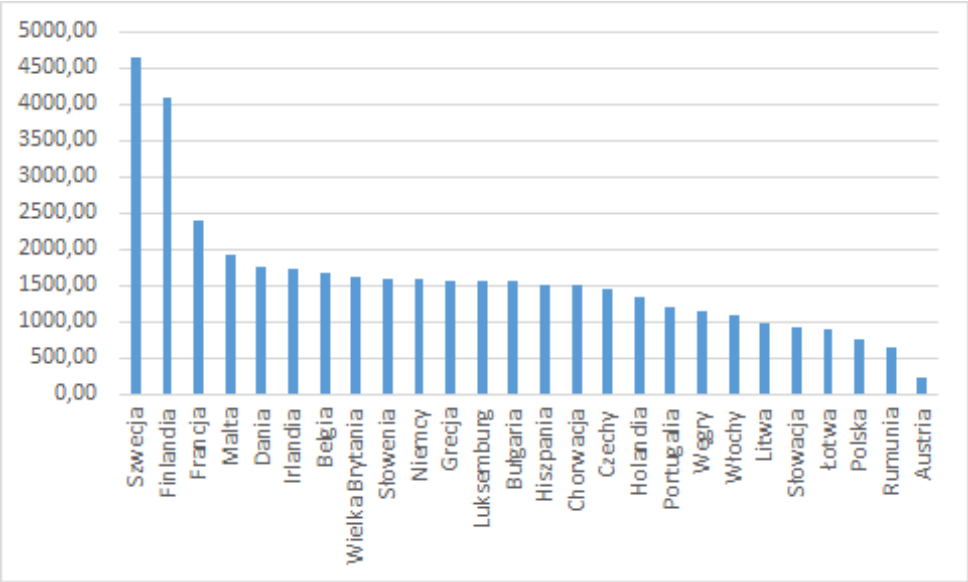


Rysunek 1.1. Zużycie energii elektrycznej w wybranych krajach Unii Europejskiej w roku 2016 (w GWh)

Źródło: Opracowanie własne na podstawie danych z Eurostat (Eurostat, 2019).

Obserwowany w ostatnich dziesięcioleciach rozwój gospodarczy, rosnąca liczba ludności oraz wielkość energii przypadająca na jednego mieszkańca spowodowały ogromny wzrost zapotrzebowania na energię elektryczną. Zwiększanie się konsumpcji prądu jest wywołane również innymi różnorodnymi czynnikami wewnętrznymi jak i zewnętrznymi, m.in. klimatem, poziomem technicznym i technologicznym, modelem życia społecznego, wskaźnikiem urbanizacji oraz strukturą gospodarki (Malko, 2006).

Europejska polityka energetyczna to dynamicznie rozwijająca się dziedzina, przemianom podlegają zarówno środki jak i jej cele. Obecnie do najistotniejszych zamierzeń polityki energetycznej w Unii Europejskiej należy rozpowszechnianie i rozwój inteligentnych sieci elektro-



Rysunek 1.2. Zużycie energii elektrycznej per capita w wybranych krajach Unii Europejskiej w roku 2016 (w kWh/os)

Źródło: Opracowanie własne na podstawie danych z Eurostat (Eurostat, 2019).

energetycznych i czujników zbierających informacje o zużyciu prądu. Z tymi innowacyjnymi przedsięwzięciami powiązane będą wysokości regulowanych taryf dystrybucyjnych z jakością zasilania oraz poziomem inwestycji w nowe technologie. Komisja Europejska wydała regulacje, które zakładają wyposażenie w inteligentne czujniki pomiaru, 80% gospodarstw domowych do roku 2020 (Michalski, 2004).

1.3 Charakterystyka rynku energii elektrycznej w Wielkiej Brytanii

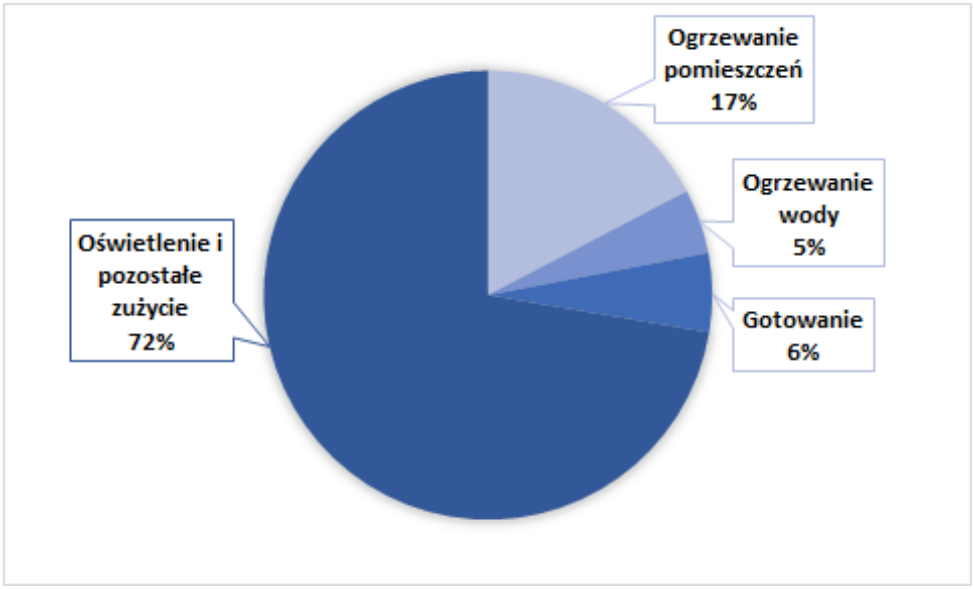
Wielka Brytania jest uznawana za prekursora reform rynku energii elektrycznej w Europie. Główną i jedną z pierwszych modyfikacji wprowadzonych w tym kraju była koncepcja liberalizacji sektorów energetycznych, uruchomienie nowych mechanizmów rynkowych oraz zapewnienie bezpiecznych dostaw energii elektrycznej. Metamorfozy przeprowadzone w Wielkiej Brytanii stanowiły wzorzec dla Komisji Europejskiej w przygotowywaniu rozporządzeń dla pozostałych krajów wspólnoty (Szablewski, 2012).

W Wielkiej Brytanii widoczna jest różnorodna struktura wykorzystania źródeł energii pierwotnej. Największy udział w niej, przypada dla gazu ziemnego i ropy naftowej. W ubiegłych

dekadach największy odsetek stanowił węgiel kamienny, jednak zaczęto znacznie ograniczać jego wykorzystanie na rzecz pozostałych zasobów oraz energii jądrowej. Obecny odsetek zużycia poszczególnych dóbr wynika przede wszystkim z zasobów naturalnych Wielkiej Brytanii i odkrycia złóż gazu ziemnego i ropy naftowej (Kaliski, Frączek & Szurlej, 2011).

Planowane jest zwiększenie znaczenia energii jądrowej w strukturze wytwarzania energii elektrycznej w Wielkiej Brytanii. Obecnie w kraju tym, działa 18 reaktorów jądrowych, z czego do 2023r. 17 z nich zakończy swoją działalność, a ostatni zostanie wyłączony w 2035r. Dla zastąpienia starych reaktorów, zbudowane zostaną nowe. Przyczyni się to do poprawy bezpieczeństwa energetycznego kraju oraz realizacji założeń o tym, że energia jądrowa będzie w najbliższych dziesięcioleciach odgrywać fundamentalną rolę w strukturze wytwarzania energii elektrycznej. W interesie publicznym Wielkiej Brytanii leży więc pozwolenie przedsiębiorcom i przedsiębiorstwom na stworzenie nowoczesnych reaktorów jądrowych (Kaliski & Frączek, 2012).

Na rysunku 1.3 przedstawiona jest struktura zużycia energii elektrycznej przez gospodarstwa domowe w Wielkiej Brytanii w roku 2016. Największa ilość zużywanego prądu jest wykorzystywana do oświetlenia mieszkań i zasilania sprzętów RTV i AGD (oprócz służących do gotowania) znajdujących się w domostwach.



Rysunek 1.3. Struktura zużycia energii elektrycznej w Wielkiej Brytanii w roku 2016

Źródło: Opracowanie własne na podstawie danych z Eurostat (Eurostat, 2019).

Wielka Brytania przedstawiła również przejrzystą ścieżkę dla wprowadzenia i rozwoju inteli-

gentnych sieci elektroenergetycznych, aktywnie popiera ich powstawanie także w pozostałych państwach Unii Europejskiej. W ramach realizacji tych założeń, rząd Wielkiej Brytanii, nakazał obowiązkowo zamontować inteligentne liczniki energii elektrycznej we wszystkich przedsiębiorstwach do końca 2014r. Natomiast instalacja liczników w gospodarstwach domowych na terenie całego kraju ma potrwać do końca 2019r. Za przebieg tego procesu, w przeciwieństwie do pozostałych krajów Unii Europejskiej, odpowiedzialni są główni dostawcy energii elektrycznej, a nie jak w przypadku reszty państw – operatorzy systemów dystrybucyjnych. Zamontowane urządzenia pozwolą m.in. na dokładniejsze pomiary i bieżące śledzenie ilości zużywanej energii elektrycznej. Umożliwią, zamiast zryczałtowanych rachunków, określenie dokładnych kosztów wykorzystanych kWh i stosowanie przez konsumentów przedpłat (Michalski, 2013).

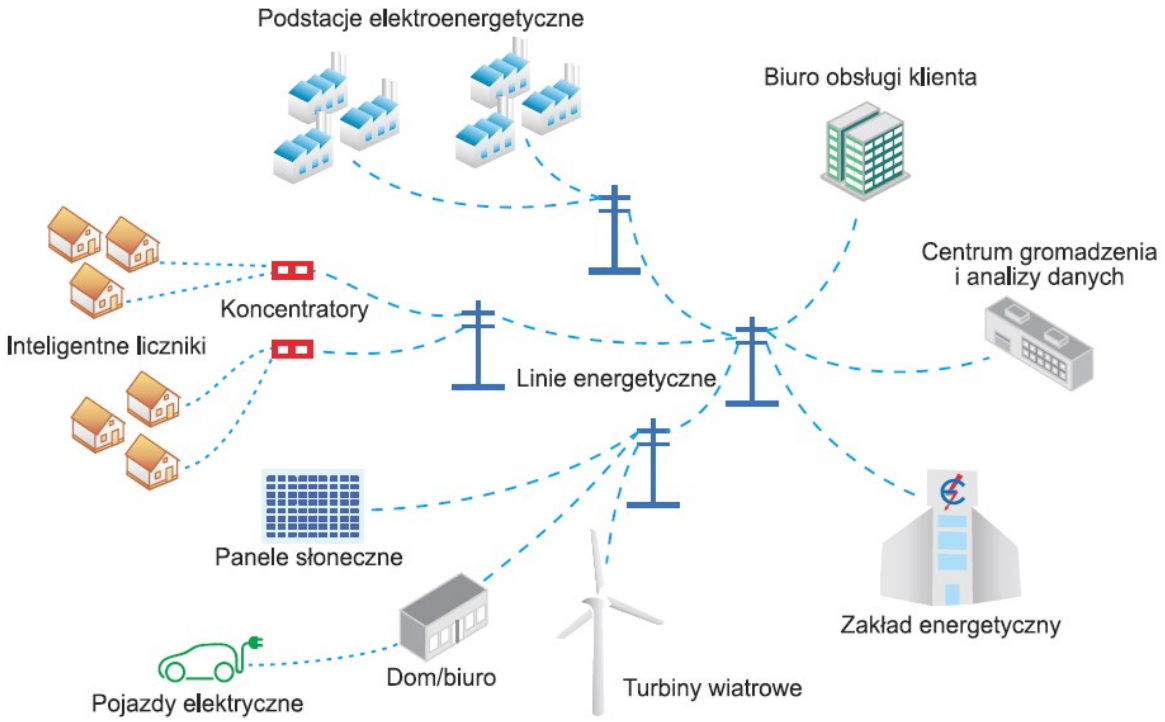
1.4 Rozwój inteligentnych sieci elektroenergetycznych

Power Line Communication (PLC) to technologia, która pozwala na przesyłanie danych przy wykorzystaniu linii sieci energetycznej. Historia początków zastosowania PLC sięga minionego wieku, kiedy to zaczęto wykorzystywać ją jako medium transmisyjne, głównie do przesyłania sygnałów zarządzających pracą systemu elektroenergetycznego i jego odbiorników. Pierwotnie było to połączenie jednokierunkowe, służące, np. do włączania i wyłączania ulicznego oświetlenia. Dopiero z upływem czasu wprowadzono również łączność dwukierunkową, co pozwoliło na znaczne rozszerzenie wykorzystania sieci PLC i przyczyniło się do powstania wielu aplikacji oraz inteligentnych sieci elektroenergetycznych.

Sieci PLC można podzielić na dwie podstawowe grupy: wąskopasmowe i szerokopasmowe. Obie sieci różnią się od siebie przede wszystkim częstotliwością transmisji, prędkością przesyłu danych oraz odległością, na jaką sygnał zostaje wysłany. Innym sposobem, w jaki można pogrupować PLC jest rodzaj użytej do transmisji danych sieci elektroenergetycznej, wyróżniamy tu: łączność za pośrednictwem sieci prądu stałego bądź sieci prądu przemiennego. Pierwsza z nich znalazła zastosowanie w wielu aplikacjach, np. jako sieć pokładowa w samochodach, pociągach i samolotach. Jednak to sieć z medium komunikacyjnym w postaci prądu przemiennego jest obecnie szybciej rozwijającą się.

Za największą zaletę sieci PLC uznaje się brak konieczności budowania nowej infrastruktury sieciowej i tworzenia dodatkowego okablowania. Przesyłanie informacji jest możliwe za pomocą istniejących już połączeń zaplecza energetycznego. Bezpośrednio wynika z tego kolejna

zaleta – sieci PLC można wykorzystywać również na terenach o słabo rozwiniętej infrastrukturze, czyli głównie na obszarach wiejskich. Dzięki temu wszystkiemu możliwe jest znaczne zaoszczędzenie na niewymaganych kosztach budowy sieci, obejmujących m.in. wydatki na samą budowę jak i utrzymanie i konserwację sprzętu. Szerokopasmowe sieci PLC są również bardzo dobrym rozwiązaniem z zakresu budowania sieci zapewniającej dostęp do internetu, transmitującej różnorodne dane multimedialne w obrębie budynków, a także w systemach domowej automatyki. Spośród wymienianych przez krytyków wad sieci PLC najczęściej zwracana jest uwaga na problem zabezpieczenia przesyłanych danych oraz ochronę przed dostępem osób nieupoważnionych. Pomimo tych wad, zainteresowanie i zastosowanie sieciami PLC ostatnimi czasy znacząco wzrasta. Przewiduje się, że sieci wąskopasmowe zostaną dominującą technologią komunikacyjną i będą wkrótce powszechnie wykorzystywane w różnych segmentach dynamicznie rozwijających się inteligentnych sieci energetycznych zwanymi smart grid (Jaworowska, 2012)



Rysunek 1.4. Przykład inteligentnej sieci elektroenergetycznej

Źródło: Opracowanie na podstawie
https://elektronikab2b.pl/technika/16187-plc-stadardem-przyszlosci#.Vw062kcmA4

Na rysunku 1.4 ukazana jest idea inteligentnej sieci elektroenergetycznej przyjmująca założenie dwukierunkowej komunikacji pomiędzy zakładem energetycznym a konsumentami oraz integrację rozproszonych źródeł energii

Wprowadzenie technologii inteligentnych sieci energetycznych ma na celu przede wszystkim: wzmocnić gospodarkę elektroenergetyczną, ulepszyć działanie sieci przemysłowej, usprawnić zarządzanie ilością wytwarzanej energii, zmienić efektywność gospodarowania naturalnymi zasobami i racjonalizować ilość zużywanych paliw i energii. Dzięki zastosowaniu smart grid można uzyskać połączenie energii produkowanej z odnawialnych źródeł energii, głównie słonecznej, wiatrowej i biomasy, które mogą posłużyć do zasilenia publicznej sieci energetycznej. Smart grid jest w stanie samodzielnie przekierować energię w chwili, gdy jej przesył zostanie zakłócony w wyniku niespodziewanej awarii. Przykładem takiej sytuacji jest nagła kilkunastogodzinna przerwa w dostawie energii w kwietniu 2008 roku w Szczecinie, która była spowodowana obfitymi opadami deszczu. Jednym z zadań inteligentnych sieci elektroenergetycznych w takiej sytuacji jest uporanie się ze zmniejszeniem liczby i długości przerw w zasilaniu.

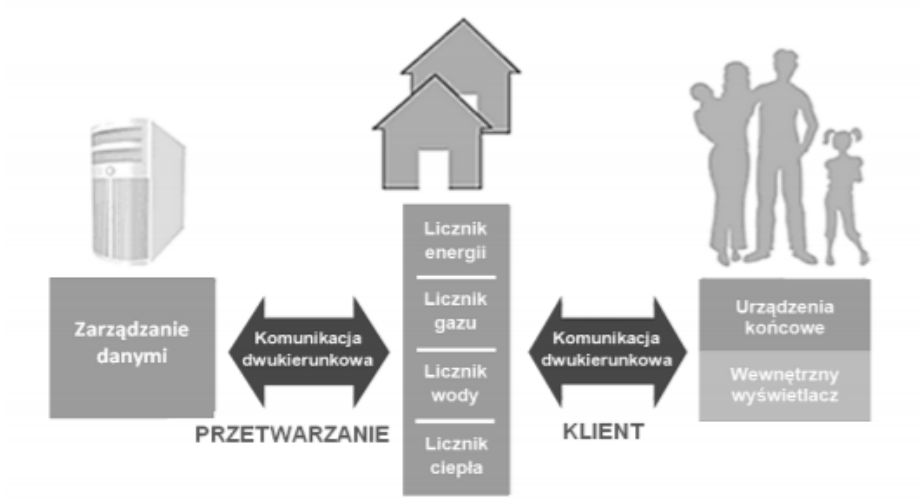
Smart grid umożliwia wspieranie energii wytwarzanej w tradycyjnych elektrowniach węglowych poprzez dostarczanie energii ze źródeł odnawialnych w taki sposób, że przekłada w czasie zasilanie urządzeń mniej istotnych aż do chwili, gdy do odbiorcy zostanie dostarczona energia pochodzące ze źródeł odnawialnych. W momentach silnych obciążeń sieciowych może wyłączyć urządzenia, których zasilanie jest obecnie najmniej istotne, natomiast poza godzinami, w których zużycie jest największe naładować akumulatory czy ogrzać wodę.

Wykorzystując dane, które są generowane w trakcie pracy inteligentnych sieci, elektrownie i przedsiębiorstwa z branży energetycznej mogą prognozować zapotrzebowanie na energię, określać w jakich momentach wzrasta, a w jakich maleje zapotrzebowanie na energię. Na podstawie wniosków wyciągniętych z tych danych, zakłady są w stanie informować klientów o aktualnej cenie energii elektrycznej, po to by skłonić swoich klientów do racjonalnego wykorzystania energii elektrycznej (Zajkowski & Borowska, 2016).

Smart grid może być wykorzystywana również do komunikacji pomiędzy urządzeniami znajdującymi się w mieszkaniach, tj. pralką, zmywarką, piekarnikiem a także regulowaniem oświetlenia i temperatury, na podstawie których jest w stanie dobrać optymalne i zgodne z preferencjami klienta poziomy ogrzewania lub klimatyzacji. Pojazdy elektryczne i hybrydowe, które korzystają z inteligentnej sieci elektroenergetycznej mogą pobierać i magazynować energię w czasie ładowania a także dostarczać ją z powrotem w momentach większego zapotrzebowania na energię elektryczną (Jaworowska, 2012).

Inteligentna sieć elektroenergetyczna wraz z inteligentnymi licznikami, według (Kubiak & Urbaniak, 2009) powinna zapewniać:

- precyzyjny pomiar zużycia energii elektrycznej
- infrastrukturę gwarantującą transmisję danych
- dopasowane do wielkości danych środowisko informatyczne
- system fakturowania zindywidualizowany dla poszczególnych klientów



Rysunek 1.5. Komunikacja za pomocą inteligentnej sieci elektroenergetycznej

Źródło: Opracowanie na podstawie https://rynek-gazu.cire.pl/pliki/2/systemy_monitorowania.pdf

1.5 Czujniki zbierające informacje o zużyciu prądu

Inteligentny licznik (ang. *Smart meter*) to zaawansowany licznik, który mierzy zużycie energii elektrycznej dostarczając przy tym większą ilość informacji w porównaniu z konwencjonalnym, wszystkim znanym licznikiem. Instalacja inteligentnych liczników, które są nieprzerwanie połączone z inteligentną siecią elektroenergetyczną wymaga wdrożenia różnych technik i oprogramowania. Jest to uzależnione również od funkcji, które ma spełniać inteligentny czujnik oraz od bieżącej sytuacji i konkretnego gospodarstwa domowego. Samo projektowanie inteligentnego czujnika zależy od wymagań dostawcy energii, przedsiębiorstwa oraz docelowego klienta. Czujnik ten można zintegrować z całą gamą innych funkcji i technologii. Wdrożenie inteligentnych czujników wymaga odpowiedniego pomiaru, przygotowania i przede wszystkim zachowania wysokich standardów bezpieczeństwa inteligentnej komunikacji sieciowej (Depuru, Wang, Devabhaktuni & Gudi, 2011).

Inteligentne liczniki energii elektrycznej są w coraz większym stopniu wykorzystywane w prywatnych domach na całym świecie. Wynika to przede wszystkim z wygody ich użytkowania dla właścicieli lokali oraz decyzji rządu mających na celu dążenie do realizacji celów oszczędności energii. W konsekwencji powstała stale rosnąca sieć komunikacyjna, składająca się z milionów lokalnych liczników, która generuje mnóstwo korzyści, zarówno dla producentów i dystrybutorów energii, organów administracyjnych jak i dla samych właścicieli gospodarstw domowych, w których to inteligentne liczniki zostały zainstalowane. Za sprawą tych urządzeń uległy znacznemu uproszczeniu transakcje z zakresu odczytu liczników, rozliczeń okresowych i zarządzania dostawami energii. Konsument uzyskał wgląd w dane o ilości energii zużywanej przez poszczególne urządzenia w jego mieszkaniu, które są w danym momencie podłączone do sieci. Na podstawie przesyłanych przez liczniki informacji dostawca energii elektrycznej jest w stanie automatycznie rejestrować wyniki pomiarów zużycia energii, wykrywać ewentualne awarie i przerwy w dostawach prądu, kontrolować aktualne obciążenie sieci i dostosowywać taryfy indywidualnie do potrzeb konsumentów, np. stosując zmienną cenę za jednostkę energii o różnych porach dnia i nocy. Z uwagi na to, że wraz z wprowadzeniem czujników rozrosła się ich cyfrowa sieć komunikacyjna, otworzyło to zupełnie nowy obszar dla aplikacji mobilnych, które teraz mogą być dostarczane na rynek. Dlatego oczywiste jest, że same liczniki, zwiększając potencjał i możliwości rynku, stały się ważnym elementem rozbudowanej struktury, która jest dostępna nie tylko dla przedsiębiorstw, ale też dla końcowych użytkowników energii elektrycznej, zapewniając im m.in. usługi z zakresu wyświetlania i zarządzania aktualnym zużyciem energii w mieszkaniu, optymalne ogrzewanie, klimatyzacje, oświetlenie, które winno być zintegrowane z innymi sieciami automatyki domowej. Do tej sieci mogłyby dołączyć również inne liczniki, np. gazu ciepła i wody tworząc podobne rozwiązania i zwiększając skuteczność usług (Benzi, Anglani, Bassi & Frosini, 2011).

Zgodnie z Michalski (2013) korzyści wynikające ze stosowania inteligentnych sieci i liczników zaowocują przede wszystkim:

- zwiększeniem się bezpieczeństwa energetycznego,
- pomocą w zwalczaniu niepokojących skutków zmian klimatu i działalności człowieka,
- wsparciem dla wzrostu gospodarczego krajów,
- zredukowaniem popytu na energię w szczytowych godzinach,
- dostarczaniem w sposób niezakłócony i nieprzerwany energii elektrycznej.

1.6 Podsumowanie

W rozdziale zostało wyszczególnione jak ważna, zarówno dla przedsiębiorstw jak i gospodarstw domowych, jest energia elektryczna. Nie sposób wyobrazić sobie, jak świat mógłby funkcjonować bez jej istnienia. Energia elektryczna jest obecnie wykorzystywana w niemalże każdym aspekcie życia. Przedstawione zostały początki kształtowania się rynku energii elektrycznej oraz najważniejsze przemiany jakie na nim zachodziły.

W dzisiejszych czasach, zauważalny jest bardzo szybki wzrost i rozwój technologiczny. Rynek energii elektrycznej również musi nieustannie ewaluować. Przykładem najnowszych tendencji na tym rynku jest wprowadzanie i rozwój inteligentnych sieci elektroenergetycznych oraz inteligentnych urządzeń do pomiaru zużywania prądu.

Rozwiązania te dają wiele możliwości, a także generują ogromne ilości danych, które mogą zostać wykorzystane do różnorodnych analiz. Przykładem jednego z badań, które można przeprowadzić na tych danych jest analiza skupień. Jej teoretyczne podstawy i najważniejsze metody znajdują się w poniższym rozdziale.

Rozdział 2

Teoretyczne podstawy analizy skupień

2.1 Analiza skupień i obszary jej zastosowań

Analiza skupień uznawana jest za jedną z najważniejszych części składowych statystycznej analizy danych wielowymiarowych. Aktualnie dostrzegane jest więcej zjawisk społeczno-ekonomicznych o coraz bardziej skomplikowanej strukturze, a w konsekwencji zwiększenie się liczby zakładanych baz danych oraz wzrost ich rozpiętości. W wyniku tych działań znaczenie, waga a także wymagania co do analizy skupień konsekwentnie wzrastają (Migdał-Najman & Najman, 2013).

Według Krzysko, Wołyński, Górecki i Skorzybut (2008) analiza skupień jest narzędziem analizy danych służącym do grupowania n obiektów, opisanych za pomocą wektora p cech, w k niepustych, rozłącznych i możliwie jednorodnych grup – skupień.

Analiza skupień polega na rozdzieleniu obserwacji, rekordów lub przypadków na zbiory o podobnych obiektach. Ogólny problem badawczy w tej analizie polega na takim podzieleniu na grupy całego zbioru rekordów, by obiekty znajdujące się wewnątrz każdej z klas były do siebie pod względem analizowanej cechy maksymalnie podobne, a ich podobieństwo do rekordów z innych grup było jak najmniejsze (Larose & Wilbik, 2013).

Nadrzędnym celem w tej analizie jest wyszukanie w zbiorze danych naturalnych skupień, które kolejno można wykorzystywać do formułowania wniosków i różnokierunkowo interpretować (Krzysko i in., 2008).

Przeprowadzając analizę skupień musimy znaleźć odpowiedź m.in. na pytania (Krzysko i in., 2008):

- Jak mierzyć podobieństwo, a jak niepodobieństwo?

- W jaki sposób standaryzować lub znormalizować zmienne ilościowe?
- Do ilu grup przydzielić obiekty?

Etapy przeprowadzania typowej analizy skupień według Walesiak i Gatnar (2009) są następujące:

1. Wybranie obiektów i zmiennych.
2. Dobranie odpowiedniej metody normalizacji wartości zmiennych.
3. Wybranie miary odległości (jednak etap ten pomija się, gdy analiza skupień bazuje bezpośrednio na macierzy danych).
4. Wybranie formuły klasyfikacji.
5. Ustalenie ilości klas.
6. Ocenienie wyników klasyfikacji.
7. Profilowanie klas i zinterpretowanie wyników.

Przeprowadzając analizę skupień musimy sprostać wielu problemom:

1. Wraz ze zwiększającą się wielkością klasyfikowanych obiektów, liczba dopuszczalnych podziałów zbioru n obiektów staje się ogromna. W związku z tym, z perspektywy pewnego kryterium, przeanalizowanie wszystkich możliwych podziałów zbioru n obiektów i wybór na tej podstawie najbardziej odpowiedniego z nich, nie jest możliwy do zrealizowania dla dużych liczebności zbioru obiektów. W takiej sytuacji niezbędne staje się zastosowanie różnorodnych, a przede wszystkim efektywnych algorytmów i metod klasyfikacji.
2. Drugi problem dotyczy liczby zmiennych, które opisują badane obiekty. Biorąc pod uwagę zależność, że dla jednej zmiennej otrzymujemy rozmieszczenie obiektów na prostej, następnie dla dwóch zmiennych uzyskujemy uporządkowanie obiektów na płaszczyźnie. W obydwu wymienionych wyżej przypadkach możliwe jest zwizualizowanie rozmieszczenia obiektów. Problem pojawia się, gdy chcemy uwzględnić więcej niż trzy zmienne. Wtedy do rozwiązania problemu potrzebujemy adekwatnych metod i algorytmów klasyfikacji.

3. Kolejnym elementem przesądzającym o skali trudności, jest rozmieszczenie elementów w przestrzeni klasyfikacji oraz brak powszechnie przyjętej definicji klasy, pomimo, że w literaturze można znaleźć wiele jej zastosowań w różnorodnych przypadkach. Z uwagi na to, że nadrzędnym celem analizy skupień jest badanie podobieństwa i odrębności obiektów i ich zbiorów, należy podzielić wszystkie obserwacje na klasy, tak aby wyodrębnione grupy spełniały przede wszystkim dwa kryteria: obiekty z tych samych klas powinny być maksymalnie spójne, a by obiekty znajdujące się w różnych klasach powinny być do siebie maksymalnie niepodobne.

4. Czwartym postulatem przeważającym o trudności przeprowadzenia analizy skupień jest brak ujednoliconej i szeroko akceptowanej teorii klasyfikacji.

Przykłady zastosowań analizy skupień podaje Larose i Wilbik (2013):

- wyszukanie grupy potencjalnych konsumentów pewnego produktu z niszy rynkowej, wytworzonego przez niewielkie przedsiębiorstwo z małym funduszem reklamowym,
- rozplanowanie zachowań finansowych na pozytywne i niepewne w ramach kontroli oblicze,
- grupowanie aktywności genów, kiedy bardzo duża ich ilość okazuje zbliżone zachowanie,
- zredukowanie liczby wymiarów w sytuacji, gdy zbiór danych jest charakteryzowany przez dużą ilość cech,
- segmentacja obrazu (w algorytmach rozpoznawania obrazu),
- segmentacja bazy klientów,
- analiza koszyka zakupów.

2.2 Metody grupowania w analizie skupień

2.2.1 Metody hierarchiczne

Spośród wielu metod grupowania w analizie skupień wyróżnić można przede wszystkim metody hierarchiczne. Są one najprostszą i zarazem najczęściej używaną metodą (Krzysko i in., 2008).

W procedurze tej tworzona jest struktura drzewiasta (dendrogram) poprzez rekurencyjne połączenie (metody aglomeracyjne) lub podzielenie (metody deglomeracyjne) istniejących grup. U podstawy wszystkich algorytmów należących do tej metody jest adekwatne określenie tzw. miary niepodobieństwa obiektów (Larose & Wilbik, 2013).

Znaczna większość metod hierarchicznych nie wymaga wielu początkowych założeń dotyczących zbioru, który chcemy przeanalizować. Najczęściej obligatoryjne jest jedynie, by określona została funkcja odległości pomiędzy punktami, za pomocą której można już wstępnie uwzględnić efekt analizy danych, np. ważenie lub ekstrakcję cech, a także odległości między skupieniami. Algorytmy w tej metodzie, w zależności od doboru odmienności mogą skupiać uwagę na lokalnej „gęstości" rozkładu danych jak i ich podobieństwie na globalnym poziomie (Cena, 2018).

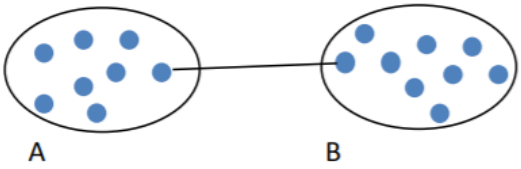
Według Larose i Wilbik (2013) idea metody aglomeracyjnej zakłada wstępnie, że każda obserwacja jest grupą składającą się wyłącznie z pojedynczego elementu, zatem ilość tych skupień wynosi n. W kolejnym kroku dwie z tych grup (skupienia) łączą się ze sobą tworząc jedną nową grupę, otrzymujemy więc $n-1$ skupień. Dzięki temu w każdym następnym ruchu, redukowana jest o jeden liczba skupień w zbiorze danych. Zatrzymanie następuje, gdy wszystkie obserwacje zostaną przydzielone do dużej pojedynczej grupy.

Graficzną ilustracją przebiegu tego procesu jest dendrogram, czyli binarne drzewo, którego liście odzwierciedlają pojedyncze obiekty, a węzły – skupienia. Liście drzewa są rozmieszczone na poziomie zerowym, natomiast węzły na poziomie adekwatnym do miary niepodobieństwa pomiędzy skupieniami wyznaczanymi przez węzły potomne (Krzysko i in., 2008).

Algorytm metody aglomeracyjnej oprócz miary niepodobieństwa pomiędzy obiektami wykorzystuje także metody wiązania skupień. Do najczęściej wykorzystywanych metod jej określania należą (Larose & Wilbik, 2013):

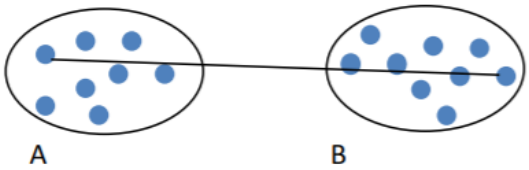
1. **Metoda pojedynczego wiązania (najbliższego sąsiedztwa)** - miara niepodobieństwa między dwoma skupieniami jest określana jako najmniejsza miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień. Zastosowanie tego typu odległości prowadzi do tworzenia wydłużonych skupień, tzw. łańcuchów.

2. **Metoda pełnego wiązania (najdalszego sąsiedztwa)** - miara niepodobieństwa pomiędzy dwoma skupieniami jest określana jako największa miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień. Zastosowanie tego typu odległości prowadzi do tworzenia zwartych skupień o małej średnicy



Rysunek 2.1. Metoda pojedynczego wiązania

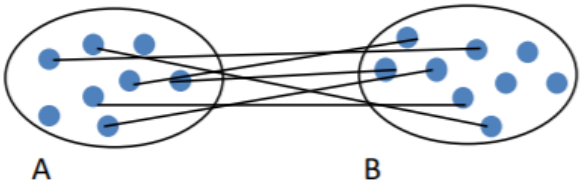
Źródło: Opracowanie na podstawie Nowak-Brzezińska (2015)



Rysunek 2.2. Metoda pojedynczego wiązania

Źródło: Opracowanie na podstawie Nowak-Brzezińska (2015)

3. **Metoda średniego wiązania** - miara niepodobieństwa między dwoma skupieniami jest określana jako średnia miara niepodobieństwa między wszystkimi parami obiektów należących do różnych skupień. Miara te jest kompromisem między metodami pojedynczego i pełnego wiązania.



Rysunek 2.3. Metoda pojedynczego wiązania

Źródło: Opracowanie na podstawie Nowak-Brzezińska (2015)

Największymi zaletami hierarchicznych metod aglomeracyjnych są (Walesiak & Gatnar, 2009):

- działanie według jednej procedury (zwanej potocznie centralną procedurą aglomeracyjną)

- przedstawienie wyników w postaci ciągu klasyfikacji (daje to szanse na kontrolowanie całego procesu klasyfikacji)
- możliwość graficznego przedstawianie rezultatów klasyfikacji – w postaci dendrogramu wskazującego kolejność połączeń pomiędzy klasami (pozwala to m.in. na ocenienie jak rozmieszczone są zarówno klasy jak i obiekty w nich usytuowane)

Z kolei metoda rozdzielająca rozpoczyna swoje działanie z wszystkimi rekordami należącymi do jednej dużej grupy. W kolejnych krokach najbardziej różniące się elementy zostają przydzielane rekurencyjnie do osobnych grup. Mechanizm działa tak długo, aż każdy z rekordów będzie reprezentował osobną grupę (Larose & Wilbik, 2013).

2.2.2 Metoda k–średnich

Według Larose i Wilbik (2013) najbardziej popularnym, niehierarchicznym algorytmem analizy skupień jest algorytm k–średnich. Zgodnie z nim przyporządkowanie n obiektów do k liczby skupień, powinno odbywać się niezależnie dla wszystkich wartości k, nie opierając się przy tym na wyznaczanych poprzednio mniejszych lub większych skupieniach. Głównym celem metody k–średnich jest takie rozmieszczenie obiektów do poszczególnych klas, które pozwoli na zminimalizowanie zmienności wewnątrz powstałych skupień oraz zmaksymalizowanie zmienności pomiędzy sąsiednimi skupieniami.

Metoda ta ma charakter iteracyjny, co w tym przypadku oznacza bazowanie na założeniu, że na wejściu znany jest wstępny podział zbioru obiektów na k klas. W metodzie tej każda z klas reprezentowana jest przez swój środek ciężkości. Wszystkie znane metody k-średnich różnią się przede wszystkim, początkowym sposobem wybrania tychże środków ciężkości w poszczególnych klasach, ich sposobem obliczania, czy zastosowaną formułą odległości (Walesiak & Gatnar, 2009).

Metoda k-średnich działa według poniższego schematu (Walesiak & Gatnar, 2009):

1. Pierwszym krokiem jest ustalenie k klas, na które wstępnie podzielimy zbiór obiektów. Liczba klas może być uzyskana dowolną metodą klasyfikacji lub nawet przyjęta losowo. Dla każdej grupy uzyskanej z wstępnego podziała oblicza się środki ciężkości a także odległości wszystkich obiektów od tych środków.
2. W kolejnym kroku należy zmienić dopasowanie obiektów do klas o najbliższym środku ciężkości.

3. Dla każdej klasy należy obliczyć nowe środki ciężkości.

Wymienione w punktach 1-3 kroki powtarza się tak długo, aż zakończą się przesunięcia obiektów pomiędzy klasami, czyli gdy wartość funkcji kryterium przestanie wykazywać istotne zmiany.

Dla ustalonej funkcji Ck (która każdemu numerowi obiektu przyporządkowuje numer skupienia) przez $W(Ck)$ i $B(Ck)$ oznaczamy macierze zmienności odpowiednio wewnątrz i między skupieniami. Otrzymujemy więc równanie (2.1):

$$W(C_K) = \sum_{k=1}^{K} \sum_{C_K(i)=k} (x_i - \overline{x}_k)(x_i - \overline{x}_k)', \tag{2.1}$$

gdzie: $\overline{x}_k$ jest wektorem średnich k-tego stopnia, a $\overline{n}_k$ jest liczebnością k-tego skupienia, tzn (2.2):

$$\overline{x}_k = \frac{1}{n_k} \sum_{C_K(i)=k} x_i. \tag{2.2}$$

Macierz $B(Ck)$ ma postać:

$$B(C_K) = \sum_{k=1}^{K} n_k (\overline{x}_k - \overline{x})(\overline{x}_k - \overline{x})', \tag{2.3}$$

gdzie średnia ogólna obliczana jest wzorem:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{2.4}$$

Poniższa zależność, znana z analizy wariancji, opisuje związek pomiędzy tymi macierzami:

$$T = W(C_K) + B(C_K), \tag{2.5}$$

gdzie, niezależna od dokonanego na skupienia podziału, macierz zmienności całkowitej T jest dana wzorem:

$$T = \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})'. \tag{2.6}$$

Funkcja realizująca optymalny podział n obiektów na K skupień dana jest wzorem:

$$C_K^* = \min_{C_K} \operatorname{tr}\left[W\left(C_K\right)\right] = \min_{C_K} \sum_{k=1}^{K} \sum_{C_K(i)=k} \rho_2\left(x_i, \bar{x}_k\right), \tag{2.7}$$

gdzie: ρ_2 oznacza kwadrat odległości euklidesowej (Larose & Wilbik, 2013).

2.3 Metoda Warda

Metoda Warda należy do hierarchicznych metod klasyfikacji obiektów w analizie skupień. Na początku przeprowadzania analizy, gdy każdy obiekt reprezentuje swoje własne skupienie, należy zdefiniować odległości pomiędzy poszczególnymi obiektami wykorzystując wybraną miarę odległości i określić, kiedy dwa skupienia są do siebie dostatecznie podobne na tyle, by można było je połączyć. W tym celu stosuje się metodę Warda, która wykorzystuje kwadrat odległości euklidesowej do pomiaru odległości między badanymi obiektami (Foryś, 2010).

Analizę skupień metodą Warda należy rozpocząć od standaryzacji zmiennych. Działanie to ma na celu dokonanie obiektywnej oceny podobieństw, nie uwzględniając jednocześnie skali, w których poszczególne zmienne zostały wyrażone. Standaryzacja pozwala na uzyskanie macierzy podobieństwa badanych obiektów, które tworzą zbiorowość. Przeprowadza się ją według wzoru:

$$z_{ij} = \frac{x_{ij} - \bar{x}_{ij}}{s_j}, \tag{2.8}$$

gdzie: i to numer obiektu, j to numer cechy, $\bar{x}_{ij}$ – średnia arytmetyczna, s_j oznacza odchylenie standardowe.

Metoda Warda do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji. Odległości pomiędzy klastrami wyznacza różnica między sumami kwadratów odchyleń poszczególnych jednostek od środka ciężkości grup, do których należą skupienia. Podsumowując, metoda ta dąży do minimalizacji sumy kwadratów odchyleń dowolnych skupień, które mogą zostać uformowane na każdym etapie. Metoda ta traktowana jest jako bardzo efektywna, chociaż zmierza do tworzenia skupień o małej wielkości (Adamowicz & Janulewicz, 2012).

Do tworzenia skupień wykorzystuje się kwadrat odległości euklidesowej, określony wzorem:

$$d(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}, \tag{2.9}$$

gdzie: x_i oraz y_i są współrzędnymi i-tego obiektu.

2.4 Weryfikacja wyników analizy skupień

2.4.1 Indeks Calińskiego i Harabasza

Indeks Calińskiego i Harabasza jest jedną z procedur pozwalającą wyznaczyć liczbę klas w analizie skupień. Optymalną wartość K dobieramy tak, by zmaksymalizować indeks Calińskiego i Harabasza, który dany jest wzorem (Walesiak & Dudek, 2007):

$$G1(u) = \frac{\mathrm{tr}(\mathbf{B})/(u-1)}{\mathrm{tr}(\mathbf{W})/(n-u)}, \quad G1(u) \in R_+, \tag{2.10}$$

gdzie: $\mathbf{B}$ to macierz kowariancji międzyklasowej, $\mathbf{W}$ to macierz kowariancji wewnątrzklasowej, tr oznacza ślad macierzy, u określa liczbę klas $(u = 2,...,n-1)$, a n to liczba obiektów.

Na podstawie wzoru (2.10) widać, że przy pomocy indeksu Calińskiego i Harabasza nie można rozstrzygnąć czy zbiór danych powinno się w ogóle dzielić na skupienia, gdyż u musi być większe od 1.

Wyznaczenie liczby skupień, na które zbiór obiektów zostanie podzielony jest etapem, który warunkuje jakość całej analizy skupień. Znacząca większość indeksów, które wyznaczają liczbę klas ma charakter optymalizacyjny, czyli dla wybranej metody podziału obiektów ustalana jest najlepsza. W przypadku indeksu Calińskiego i Harabasza za najlepszą ilość skupień przyjmowana jest największa wartość wynikająca z przeprowadzonych obliczeń (Walesiak & Gatnar, 2009).

2.4.2 Metoda HINoV

Metoda HINoV służy do selekcji zmiennych w analizie skupień. Została zaproponowana w 1999r. przez Carmone, Kara i Maxwell. Początkowo opierała się jedynie na metodzie k—średnich i skorygowanym indeksie Randa. W swej pierwotnej wersji była zupełnie nieodporna na występowanie wśród zmiennych zanieczyszczających strukturę skupień zmiennych skorelo-

wanych jednomodalnych lub równomiernych. Wada ta została z biegiem czasu wyeliminowana, a sama metoda została rozszerzona dla innych metod klasyfikacji i danych niemetrycznych Walesiak i Dudek (2007).

W wyniku zastosowania tej procedury możliwe jest uzyskanie wykresu osypiska i zobrazowanie, które zmienne zakłócają istniejącą w układzie dwuwymiarowym strukturę klas Foryś (2010).

2.4.3 Indeks Randa

Indeks Randa pozwala ocenić zgodność dwóch podziałów zbioru na rozłączne podzbiory. Indeks ten wyznacza również liczbę skupień w zbiorze danych. Jest oparty na wielostopniowym dzieleniu zbioru danych na dwa skupienia i sprawdzaniu, czy otrzymany podział należy zachować czy pominąć. Podziały dokonywane są przy pomocy metody k–średnich na podstawie wielokrotnego losowego wyboru punktów startowych.

Niech $O = (o_1, ..., o_n)$ będzie zbiorem obiektów, a $X = (x_1, ..., x_r)$ oraz $Y = (y_1, ..., y_r)$ dwoma podziałami.

Rozważając parę różnych elementów z $O = (o_i, o_j)$ otrzymujemy cztery możliwe przypadki:

- (o_i, o_j) należą do jednego ze zbiorów w X oraz do jednego ze zbiorów w Y
- (o_i, o_j) należą do dwóch różnych zbiorów w X oraz różnych zbiorów w Y
- (o_i, o_j) należą do tego samego zbioru w X oraz różnych zbiorów w Y
- (o_i, o_j) należą do różnych zbiorów w X oraz tego samego zbioru w Y

Powstałe przypadki oznaczamy kolejno jako A, B, C i D

$$RI = \frac{A+B}{A+B+C+D} \tag{2.11}$$

Wzór (3.2) wyznacza wielkość indeksu Randa, która zawiera wartość pomiędzy 0 a 1 i może być interpretowana jako prawdopodobieństwo, że dowolnie wybrana para jest w analogiczny sposób sklasyfikowana w obu grupowaniach.

Wzór (2.12) jest skorygowanym indeksem Randa, który wprowadza poprawkę uwzględniającą prawdopodobieństwo, że dwa algorytmy grupowania zachowując się losowo, równocze-

śnie rozdzielą parę lub dołączą do jednej grupy.

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \tag{2.12}$$

Im wyższa wartość współczynnika Randa, tym większe prawdopodobieństwo, że analiza skupień została przeprowadzona prawidłowo, a ilość klas została słusznie określona (Walesiak, 2005).

2.5 Podsumowanie

Analiza skupień jest jedną z metod statystycznej analizy danych wielowymiarowych. Analiza skupień dzieli zbiór obiektów na jednorodne grupy o podobnych obiektach na podstawie analizowanych cech.

W rozdziale tym zostały ukazane jej teoretyczne podstawy i najważniejsze metody, należą do nich przede wszystkim metody hierarchiczne oraz metoda k–średnich. Oprócz tych dwóch głównych procedur omówione zostały również metody służące do ustalania liczby skupień oraz weryfikacji poprawności samej analizy.

Kolejny rozdział przedstawia badanie wykorzystujące teoretyczne podstawy analizy skupień, przeprowadzone na danych o zużyciu prądu w gospodarstwach domowych w Londynie.

Rozdział 3

Empiryczna ocena grupowania gospodarstw domowych w Londynie

3.1 Eksploracyjna analiza danych

Na potrzebę badania wybrano i wykorzystano dane pochodzące ze strony https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households. Dane zawierają informacje o ilości zużytej energii elektrycznej przez 5 567 gospodarstw domowych mieszczących się w Londynie w Wielkiej Brytanii. Dane obejmują swym zakresem okres od listopada 2011r. do lutego 2014r. Odczyty ilości zużywanych kWh w każdym z gospodarstw były wykonywane w półgodzinnych interwałach, co daje 48 rekordów każdej doby. Klienci biorący udział w badaniu zostali sklasyfikowani jako zrównoważona reprezentacja populacji mieszkańców Londynu. Dla potrzeb badania, ze względu na braki danych, z całego zbioru wybrano informacje o zużyciu energii elektrycznej dla 5 490 gospodarstw domowych w lutym 2013r. Braki danych mogą wynikać, np. z tymczasowego braku zużywania prądu w mieszkaniu lub uszkodzenia licznika energii elektrycznej.

Pierwszym krokiem w przeprowadzonym badaniu było obliczenie podstawowych statystyk opisowych, które zawiera tabela 3.1.

Minimalna ilość zużytej energii elektrycznej dla jednego gospodarstwa domowego w ciągu 30 minut wyniosła 0,001 kWh. Kwartyl pierwszy wskazuje, że 25% gospodarstw domowych w ciągu 30 minut zużywało 0,064 kWh lub mniej. Połowa badanych gospodarstw domowych w ciągu 30 minut zużywa 0,135 kWh lub mniej. W ciągu 30 minut jedno gospodarstwo domowe zużywało przeciętnie 0,2598 kWh. Kwartyl trzeci wskazuje, że 75% gospodarstw domowych

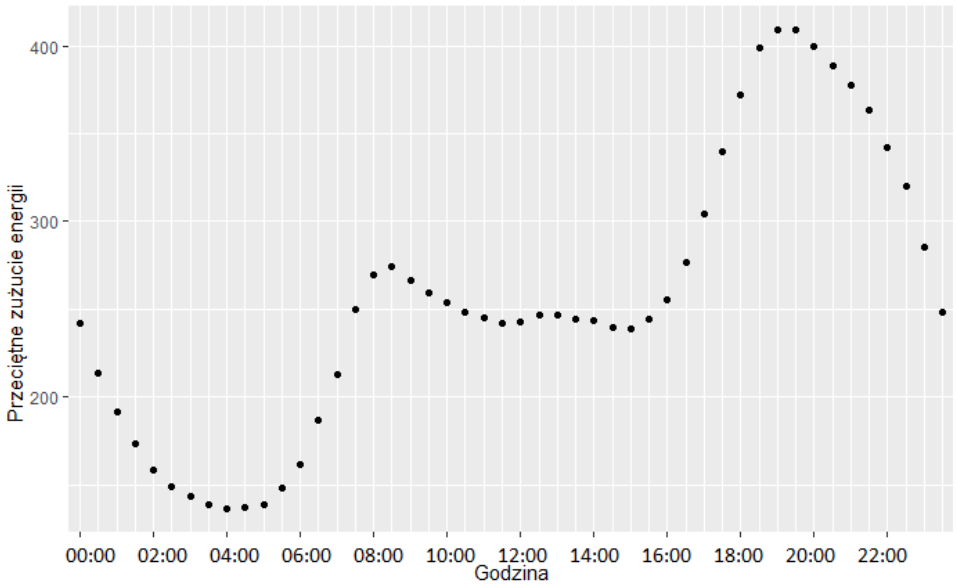
w ciągu 30 minut zużywało 0,290 kWh lub mniej. Maksymalna ilość zużytej energii elektrycznej dla jednego gospodarstwa domowego w ciągu 30 minut wyniosła 7,527 kWh.

Tabela 3.1. Podstawowe statystyki opisowe

Nr.	Statystyka opisowa	Wartość w kWh
1	Minimum	0,001
2	Kwartyl 1	0,064
3	Mediana	0,135
4	Średnia	0,2598
5	Kwartyl 3	0,290
6	Maksimum	7,527

Źródło: Opracowanie własne.

Na wykresie 3.1 ukazana została przeciętna ilość zużywanej energii elektrycznej przez gospodarstwa domowe w dni robocze w 30 minutowych interwałach. Z wykresu wynika, że największa ilość zużywanej energii elektrycznej przypadała na godzinę 19:00, a najmniejsza na godzinę 04:00. Ilość zużywanej w poszczególnych godzinach w ciągu doby, energii elektrycznej cechuje się dużą zmiennością, wynika to przede wszystkim z naturalnego trybu i planu dnia każdego człowieka.

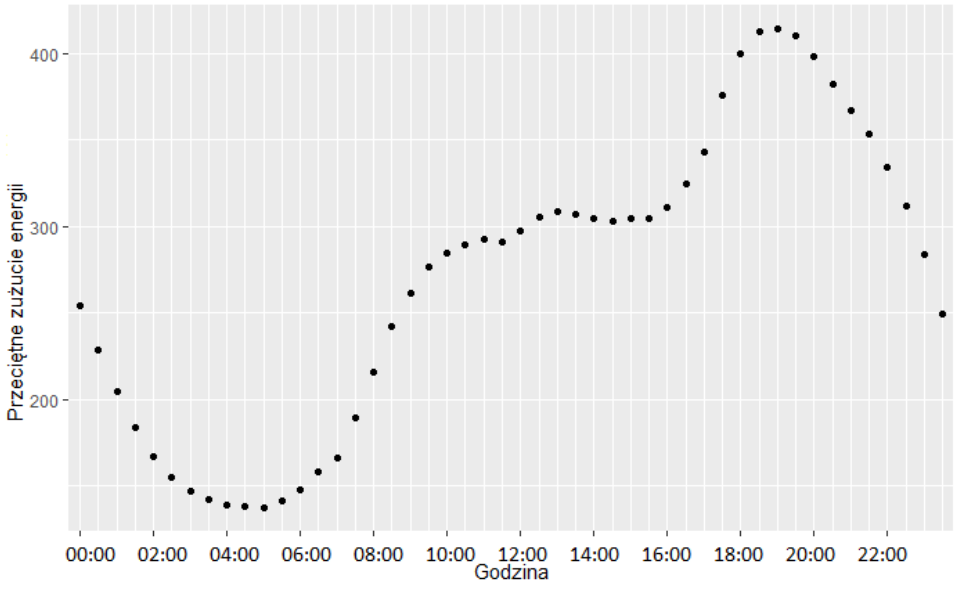


Rysunek 3.1. Przeciętna ilość zużywanej energii elektrycznej w dni robocze

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

Na wykresie 3.2 ukazana została przeciętna ilość zużywanej energii elektrycznej przez gospodarstwa domowe w soboty i niedziele w 30 minutowych interwałach. Z wykresu wynika,

że największa ilość zużywanej energii elektrycznej przypadała również na godzinę 19:00, a najmniejsza na godzinę 05:00. Ilość zużywanej w poszczególnych godzinach w ciągu doby, energii elektrycznej cechuje się dużą zmiennością, wynika to przede wszystkim z naturalnego trybu i planu dnia każdego człowieka. Można zauważyć, że zużycie energii elektrycznej w soboty i niedziele jest nieco większe od ilości zużywanej energii elektrycznej w pozostałe dni. Jest to widoczne przede wszystkim w godzinach od 08:00 do 16:00. Można uzasadnić to założeniem, że w tym czasie od poniedziałku do piątku ludzie przebywają najprawdopodobniej w pracy, natomiast weekendy spędzają w swoich mieszkaniach.



Rysunek 3.2. Przeciętna ilość zużywanej energii elektrycznej w soboty i niedziele

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

3.2 Dobór zmiennych

W celu przeprowadzenia analizy skupień i sklasyfikowania gospodarstw domowych z Londynu według ilości zużywanej energii elektrycznej utworzono jedenaście zmiennych diagnostycznych kierując się informacjami zawartymi w artykule ”Household Classification Using Smart Meter Data” (Carroll, Murphy, Hanley, Dempsey & Dunne, 2018) a także względami merytoryczno–formalnymi oraz wartościami informacyjnymi zmiennych.

- X1 – przeciętna ilość zużytej energii elektrycznej przez jedno gospodarstwo domowe

w ciągu 30 min w kWh, wyrażana wzorem (3.1):

$$E_1 = \frac{1}{l}\sum_{i=1}^{l} E_i, \tag{3.1}$$

- X2 – maksymalna ilość zużytej energii elektrycznej przez jedno gospodarstwo domowe w ciągu 30 min w kWh, wyrażana wzorem (3.2):

$$E_2 = \max\left(\{E_i\}\right), \tag{3.2}$$

gdzie $1 \leq i \leq l$

- X3 – ilość zużytej energii elektrycznej przez jedno gospodarstwo domowe w lutym 2013r., wyrażana wzorem (3.3):

$$E_3 = \sum_{i=1}^{l} E_i, \tag{3.3}$$

- X4 – przeciętne maksymalne dzienne zużycie energii elektrycznej przez jedno gospodarstwo domowe, wyrażana wzorem (3.4):

$$E_4 = \frac{1}{m}\sum_{j=1}^{m} E_j, \tag{3.4}$$

gdzie: $E_j = \max\left(\{E_i\}\right)$ oraz $1 + n(m-1) \leq i \leq nm$

- X5 – wskaźnik średniego dziennego zużycia do maksymalnego dziennego zużycia energii elektrycznej przez jedno gospodarstwo domowe, wyrażany wzorem (3.5):

$$E_5 = \frac{(1/n)\sum_{i=1}^{n} E_i}{\max\left(\{E_i, 1 \leq i \leq n\}\right)}, \tag{3.5}$$

- X6 – wariancja, wyrażana wzorem (3.6):

$$E_6 = \frac{1}{l}\sum_{i=1}^{l} \left(E_i - E_{mea\,n}\right)^2, \tag{3.6}$$

- X7 – odchylenie standardowe, wyrażane wzorem (3.7):

$$E_7 = \sqrt{E_6}, \tag{3.7}$$

- X8 – różnica pomiędzy maksymalnym a minimalnym dziennym zużyciem energii elektrycznej przez jedno gospodarstwo domowe, wyrażana wzorem (3.8):

$$E_8 = \max\left(\{E_i, 1 \leq i \leq l\}\right) - \min\left(\{E_i, 1 \leq i \leq l\}\right), \tag{3.8}$$

- X9 – różnica między trzecim a pierwszym kwartylem
- X10 – różnica między maksymalnym a średnim zużyciem energii elektrycznej przez jedno gospodarstwo domowe pomiędzy godziną 10:00 a 12:00
- X11 – różnica między maksymalnym a minimalnym zużyciem energii elektrycznej przez jedno gospodarstwo domowe pomiędzy godziną 10:00 a 12:00

gdzie:

- l – całkowita liczba półgodzinnych interwałów w lutym 2013r.
- n – całkowita liczba półgodzinnych interwałów w ciągu doby
- m – całkowita liczba dni w lutym 2013r.
- E – ilość zużytej energii w kWh

Po utworzeniu zmiennych wykonano analizę korelacji i sprawdzono, czy otrzymane cechy są ze sobą istotnie statystycznie powiązane.

Graficzną interpretacją powiązań między zmiennymi jest korelacyjny wykres rozrzutu 3.3. Wykres został wygenerowany przy użyciu funkcji `pairs` z pakietu graphics.

Cały wykres składa się ze 110 wykresów reprezentujących dwuwymiarową płaszczyznę, gdzie jedna oś odpowiada wynikom dla jednej zmiennej, natomiast druga oś odpowiada wynikom drugiej zmiennej. Wykres jest graficzną interpretacją korelacji między zmiennymi X1–X11.

Wykres ten jest dostarcza wielu informacji, dzięki niemu można uchronić się przed poważnym błędem przy przeprowadzaniu analizy skupień, czyli zakwalifikowaniu przypadków odstających. Na jego podstawie można wywnioskować m.in., że:

- zmienne X2 i X8 są ze sobą bardzo silnie dodatnio skorelowane, czyli wraz ze wzrostem maksymalnej ilości zużytej energii elektrycznej przez jedno gospodarstwo domowe w ciągu 30 min wzrasta również wartość różnicy pomiędzy maksymalnym a minimalnym dziennym zużyciem energii elektrycznej przez jedno gospodarstwo domowe

- zmienne X7 i X11 są ze sobą słabo skorelowane, czyli wraz ze wzrostem odchylenia standardowego dla ilości zużywanej energii elektrycznej wartość różnicy pomiędzy maksymalnym a minimalnym zużyciem energii elektrycznej przez jedno gospodarstwo domowe w godzinach 10:000-12:00 nie zwiększa się, między zmiennymi nie ma praktycznie żadnej zależności

Kolejnym krokiem w przeprowadzaniu badania było obliczenie dokładnych wartości współczynnika korelacji Spearmana. W tym celu użyto funkcji `cor` z pakietu stats. Do graficznej interpretacji otrzymanych wyników użyto natomiast funkcji `corrplot.mixed` z pakietu corrplot.

Na otrzymanym wykresie 3.4 ukazane są wartości wartości współczynnika korelacji Spearmana dla zmiennych X1–X11. Wszystkie wartości zawierają się w przedziale [-1;1]. Dzięki nim możemy uzyskać odpowiedź na pytania: czy istnieje powiązanie między dwoma wybranymi zmiennymi oraz czy znak współczynnika korelacji jest ujemny czy dodatni.

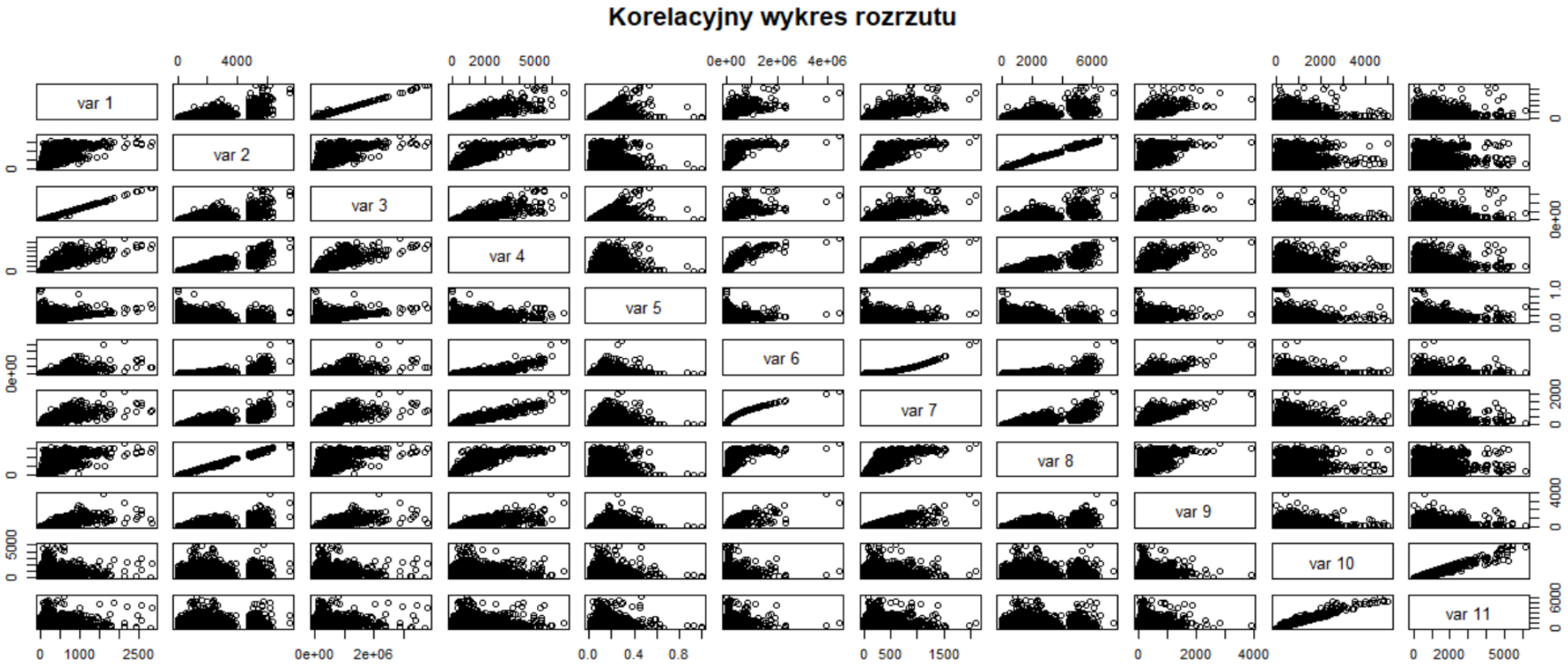
Analizując poszczególne powiązania między zmiennymi można wywnioskować, że:

- zmienne X1 i X3 są ze sobą bardzo silnie dodatnio skorelowane, czyli wraz ze wzrostem przeciętnej ilości zużywanej energii elektrycznej przez jedno gospodarstwo domowe w ciągu 30 min, rosła ilość zużywanej energii elektrycznej przez gospodarstwa domowe w lutym 2013r.

- pomiędzy zmiennymi X5 i X9 występuje umiarkowana korelacja, czyli wskaźnik średniego dziennego zużycia do maksymalnego dziennego zużycia energii elektrycznej jest umiarkowanie skorelowany z różnicą pomiędzy trzecim a pierwszym kwartylem

- zmienne X5, X10 i X11 są słabo skorelowane z wszystkimi pozostałymi zmiennymi, mogą więc zakłócać strukturę klas w analizie skupień

Przeprowadzając analizę korelacji należy pamiętać, że nie bada ona związku przyczynowo–skutkowego, tylko jedynie powiązanie między dwoma zmiennymi. Po zbadaniu korelacji nie wiemy, która cecha wpływa na którą. Udowadniamy jedynie, że wartość jednej zmiennej rośnie lub maleje w przypadku wzrostu lub spadku drugiej zmiennej.



Rysunek 3.3. Korelacyjny wykres rozrzutu

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

Rysunek 3.4. Korelacje Spearmana

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

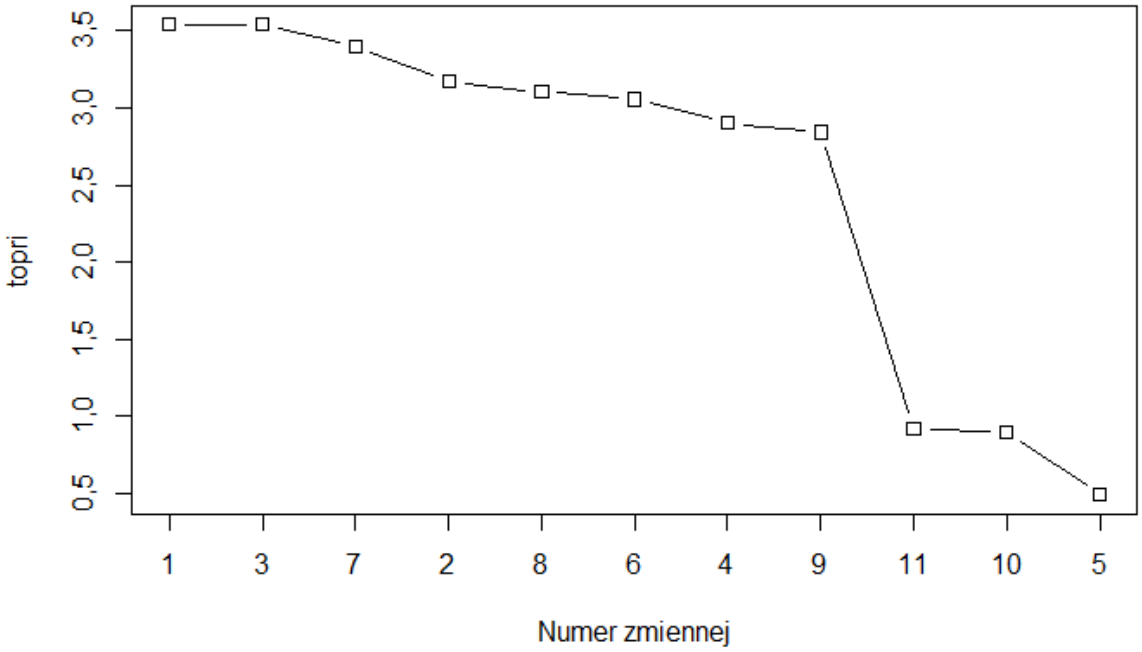
3.3 Analiza skupień

Analiza skupień została rozpoczęta od jednego z najważniejszych, a zarazem najtrudniejszych zagadnień, czyli wyboru zmiennych objaśniających. To od jakości wyboru ich zestawu zależy wiarygodność ostatecznych wyników klasyfikacji i trafność decyzji podejmowanych na ich podstawie.

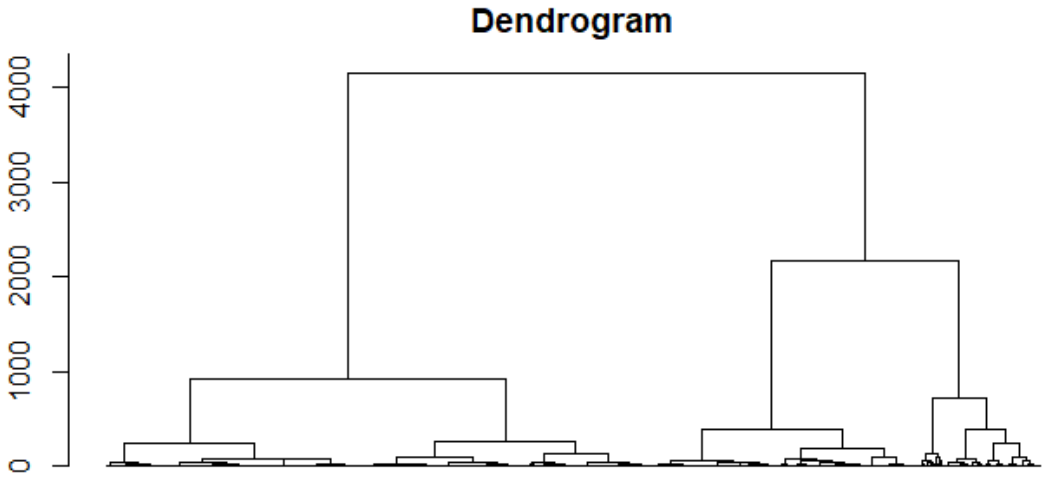
Funkcja `HINoV.Mod` (znajdująca się w pakiecie clusterSim) opiera się na metodzie k–średnich i skorygowanym indeksie Randa. Za pomocą tej funkcji wyznacza się zmienne zakłócające strukturę klas. Składowe znajdujące się na prawo od punktu kończącego osypisko reprezentują znikomą wariancję i przedstawiają w większości losowy szum. W wyniku zastosowania tej procedury otrzymano wykres osypiska 3.5, na podstawie którego dowiedziono, że zmienne 11, 10 oraz 5 należy usunąć, ponieważ zakłócają istniejącą w układzie dwuwymiarowym strukturę klas. Były to odpowiednio: różnica między maksymalnym a minimalnym zużyciem energii elektrycznej przez jedno gospodarstwo domowe pomiędzy godziną 10:00 a 12:00, różnica między maksymalnym a średnim zużyciem energii elektrycznej przez jedno gospodarstwo domowe między godziną 10:00 a 12:00 oraz wskaźnik średniego dziennego zużycia do maksymalnego dziennego zużycia energii elektrycznej przez jedno gospodarstwo domowe. Badanie korelacji wskazało, że były to jednocześnie zmienne najsłabiej powiązane z pozostałymi.

Po usunięciu zmiennych, które zakłócały liczbę klas, wykonano hierarchiczne grupowanie gospodarstw domowych metodą Warda, która wykorzystuje kwadrat odległości euklidesowej. W celu przeprowadzenia tego etapu analizy skorzystano z funkcji `hclust`, która znajduje się w pakiecie stats. W wyniku działania tej funkcji otrzymano dendrogram 3.6 ukazujący hierarchiczną strukturę podobieństw pomiędzy analizowanymi obiektami. Liczby na osi pionowej oznaczają odległości, a długości linii łączące poszczególne skupienia – najmniejsze wyszukane odległościom między nimi. Maleją one z każdym kolejnym połączeniem, ponieważ odległości wyliczane dla nowo tworzących się skupień są zawsze mniejsze niż odległości bazowe. Dendrogram umożliwia rozdzielenie od siebie jednorodnych grup gospodarstw domowych pod względem ilości zużywanej energii elektrycznej. Uzyskany wykres nie jest w pełni czytelny, ze względu na duży zbiór obserwacji. Wszystkie liście drzewa zostały wyrysowane na jednym poziomie na wykresie i ze względu na dużą liczbę badanych gospodarstw domowych nachodzą na siebie.

Kolejnym etapem w procesie przeprowadzania analizy skupień było ustalenie liczby klas, na które zbiór danych zostanie podzielony. W tym celu wykorzystano indeks Calińskiego i Harabasza. Do przeprowadzenia obliczeń użyto funkcji `index.G1` z pakietu clusterSim. Prze-



Rysunek 3.5. Wykres osypiska
Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)



Rysunek 3.6. Dendrogram przedstawiający grupowanie badanych gospodarstw domowych

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

prowadzone obliczenia nie przyniosły jednak jednoznacznej odpowiedzi na pytanie na ile grup podzielić zbiór danych.

Z uwagi na to, zastosowano zamienne rozwiązanie, które pozwoliło na uzyskanie optymalnej liczby skupień. Ostateczna liczba klas, na które zostały podzielone gospodarstwa

domowe wynika z obliczenia skorygowanego indeksu Randa. Został on policzony funkcją `replication.Mod` z pakietu clusterSim. Indeks został wyliczony dla podziału na klasy od 3 do 8. Dla podziału na 3 skupienia osiąga najwyższą wartość i świadczy o wysokiej stabilności podziału zbioru 5 490 gospodarstw domowych na 3 klasy.

Tabela 3.2. Indeks Randa

Liczba klas	Indeks Randa
3	0.8427072
4	0.7890497
5	0.6948256
6	0.6305356
7	0.8113056
8	0.7214195

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

Tabela 3.2 zawiera kolejno podział na różne liczby klas i odpowiadające im obliczone skorygowane indeksy Randa, które pozwalają ocenić zgodność dwóch podziałów zbioru na rozłączne podzbiory. Indeks Randa zwraca wartość pomiędzy 0 a 1 i może być interpretowany jako prawdopodobieństwo, że dwa algorytmy grupowania zachowując się losowo równocześnie rozdzielą parę lub dołączą do jednej grupy. Im wyższa więc wartość współczynnika Randa, tym większe prawdopodobieństwo, że analiza skupień została przeprowadzona prawidłowo.

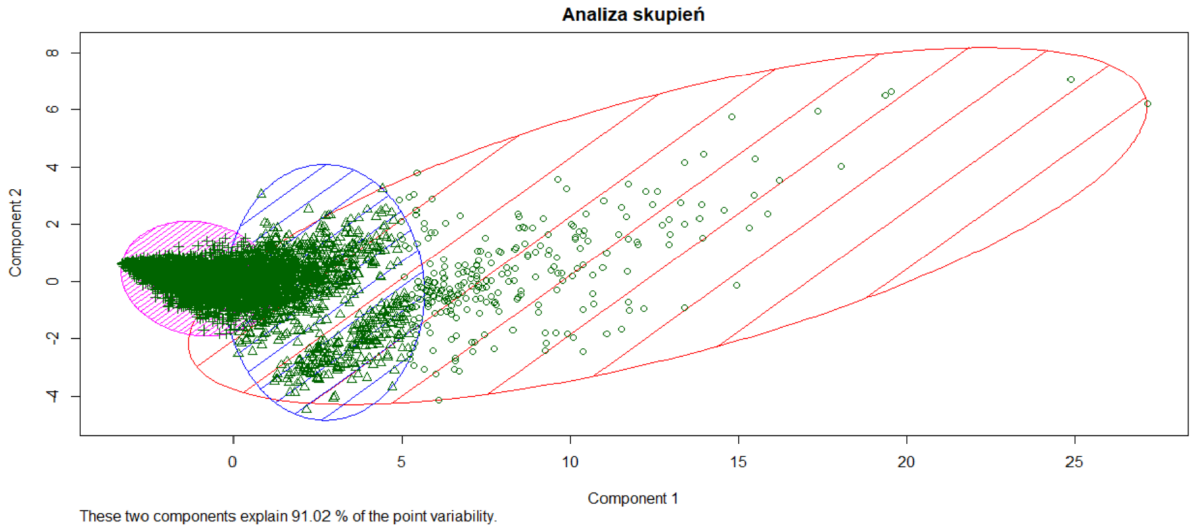
Następnym zadaniem w procesie przeprowadzania badania było utworzenie analizy skupień metodą k-średnich oraz jej graficzna prezentacja. Do wykonania tego etapu i podzielenia zbioru 5 490 gospodarstw domowych użyto funkcji `kmeans` z pakietu stats. Otrzymano podział na 3 klasy, których liczebności przedstawia poniższa tabela 3.3.

Tabela 3.3. Podział gospodarstw domowych na klasy (Datastore, 2015)

Klasa 1	Klasa 2	Klasa 3
3756	1465	269

Źródło: Opracowanie własne na podstawie danych z London Datastore

Wykres 3.7 jest graficzną reprezentacją przeprowadzonej analizy skupień metodą k-średnich. Powstał za pomocą funkcji `clusplot` z pakietu cluster. Wynikami przeprowadzonej analizy są trzy jednorodne skupienia dzielące gospodarstwa domowe ze względu na ilość zużywanej energii elektrycznej. Z uwagi na dużą ilość zmiennych otrzymany wykres nie jest w pełni czytelny, jednak można na nim dostrzec wyodrębnione w wyniku badania grupy.



Rysunek 3.7. Wizualizacja wyników analizy skupień badanych gospodarstw domowych

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

Charakterystyka skupień otrzymanych za pomocą analizy skupień wygląda następująco:

- Grupa pierwsza, największa ze względu na liczebność obiektów zawiera ich 3756, co stanowi 68,42% wszystkich gospodarstw domowych. Całkowita suma kwadratów w obrębie tej grupy stanowiła 6101,46.
- Grupa druga, kolejna ze względu na liczebność obiektów zawiera ich 1465, co stanowi 26,68% wszystkich gospodarstw domowych. Całkowita suma kwadratów w obrębie tej grupy stanowiła 5108,69.
- Grupa trzecia, najmniejsza ze względu na liczebność obiektów zawiera ich 269, co stanowi pozostałe 4,90% gospodarstw domowych. Całkowita suma kwadratów w obrębie tej grupy stanowiła 3506,99.

Tabela 3.4 przedstawia centra poszczególnych skupień dla każdej ze zmiennych.

Tabela 3.4. Centra klastrów

	X1	X2	X3	X4	X6	X7	X8	X9
1	2,9526	2,7030	2,9560	3,0218	3,1509	3,0622	2,6839	2,9945
2	0,5826	0,7587	0,5807	0,6838	0,2535	0,6784	0,7537	0,5130
3	-0,4387	-0,4895	-0,4382	-0,4831	-0,3246	-0,4839	-0,4862	-0.4146

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

W wyniku przeprowadzonej analizy skupień otrzymano trzy segmenty gospodarstw domowych. Gospodarstwa te różnią się pod względem ilości zużywanej energii elektrycznej.

Tabela 3.5 powstała dzięki wykorzystaniu funkcji `cluster.Description` z pakietu cluster. Tabela ta przedstawia przeciętne ilości energii elektrycznej zużyte przez gospodarstwa domowe w lutym 2013r.

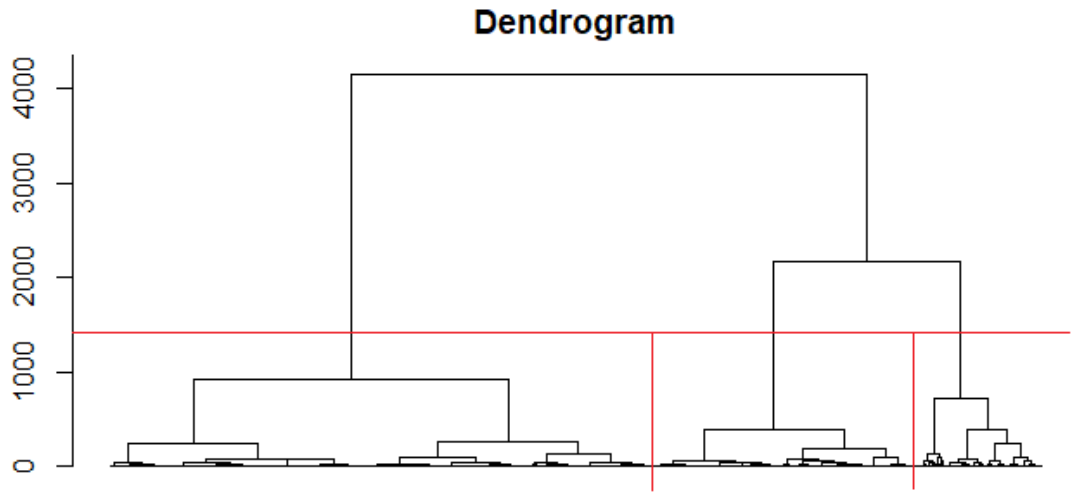
Gospodarstwa z klasy pierwszej, które stanowiły najliczniejszą grupę, zużywały najwięcej energii elektrycznej. Druga grupa okazała się pośrednia pod względem liczby mieszkań i konsumpcji energii. Ostatnia grupa była najmniejsza ze względów zarówno liczebnych jak i konsumpcyjnych. Otrzymane wyniki potwierdzają, że poddział gospodarstw domowych, na jednorodne grupy, ze względu na zróżnicowanie ilości zużywanej energii elektrycznej jest możliwy.

Tabela 3.5. Przeciętna ilość zużywanej energii elektrycznej przez jedno gospodarstwo domowe w lutym 2013r.

	Klasa 1	**Klasa 2**	**Klasa 3**
Średnia arytmetyczna	973,37	461,78	258,80

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

Na dendrogramie na rysunku 3.8 uwzględniony został podział na skupienia wynikający z rozwiązania otrzymanego po przeprowadzeniu analizy skupień metodą k-średnich. Widać na nim rozlokowanie trzech grup, do których przydzielone zostały gospodarstwa domowe.



Rysunek 3.8. Dendrogram

Źródło: Opracowanie własne na podstawie danych z London Datastore (Datastore, 2015)

Podsumowanie

Tematem pracy była analiza skupień gospodarstw domowych na podstawie informacji o ilości zużywanej energii elektrycznej przez każde z nich. Dane te zostały pozyskane dzięki rozwojowi inteligentnych sieci elektroenergetycznych oraz wykorzystaniu inteligentnych czujników do pomiaru ilości zużywanego prądu, które obecnie zyskują na znaczeniu na całym świecie.

Głównym celem pracy było uzyskanie odpowiedzi na pytanie: czy możliwe jest wyszukanie jednorodnych grup gospodarstw domowych ze względu na ilość zużywanej energii elektrycznej?

W ramach przeprowadzonych obliczeń oraz interpretacji wyników udało się odpowiedzieć na powyższe pytanie i pogrupować gospodarstwa domowe według ilości zużywanej energii elektrycznej. Otrzymane wyniki zostały przedstawione na wykresach, lecz ze względu na bardzo dużą ilość danych niektóre z nich nie są w pełni czytelne.

Na podstawie wykonanej analizy otrzymano podział gospodarstw domowych na trzy grupy. Pierwsza z nich reprezentuje gospodarstwa domowe, które zużywały najwięcej energii elektrycznej, druga grupa zawiera gospodarstwa ze średnią konsumpcją, a trzeci segment to gospodarstwa zużywające najmniej prądu. Oznacza to, że możliwe jest sklasyfikowanie gospodarstw ze względu na ilość zużywanej energii. Należy jednak pamiętać, że dane pochodzące z inteligentnego licznika energii elektrycznej i zawierające informacje wyłącznie o ilości zużywanej energii elektrycznej mają ograniczoną zdolność do rozróżniania gospodarstw domowych.

Analiza skupień jedynie wykrywa struktury zbiorów w danych, bez wyjaśniania dlaczego one występują. Otwartą kwestią i polem do przeprowadzenia kolejnych badań naukowych byłoby więc scharakteryzowanie każdej z uzyskanych w procesie klasteryzacji grup. Dzięki temu możliwe byłoby uzyskanie odpowiedzi na inne pytania, np. czy da się prognozować w których gospodarstwach i w jakich godzinach zapotrzebowanie na energię będzie maleć lub wzrastać oraz czy da się przewidzieć ilość zużywanej energii ze względu na liczbę osób stanowiących skład gospodarstwa.

Bibliografia

Adamowicz, M. & Janulewicz, P. (2012). Wykorzystanie metod wielowymiarowych w określeniu pozycji konkurencyjnej gminy na przykładzie województwa lubelskiego. *Metody ilościowe w badaniach ekonomicznych*, *13*(1), 17–28.

Benzi, F., Anglani, N., Bassi, E. & Frosini, L. (2011). Electricity smart meters interfacing the households. *IEEE Transactions on Industrial Electronics*, *58*(10), 4487–4494.

Bielecki, S. (2007). Jakość energii elektrycznej na rynku energii. *Przegląd Elektrotechniczny*, *83*.

Carroll, P., Murphy, T., Hanley, M., Dempsey, D. & Dunne, J. (2018). Household classification using smart meter data. *Journal of Official Statistics*, *34*(1), 1–25.

Cena, A. (2018). Adaptacyjne algorytmy hierarchicznej analizy skupień oparte na metodach agregacji danych.

Datastore, L. (2015). Dane z London Datastore.

Depuru, S. S. S. R., Wang, L., Devabhaktuni, V. & Gudi, N. (2011). Smart meters for power grid—Challenges, issues, advantages and status. W *2011 IEEE/PES Power Systems Conference and Exposition* (s. 1–7). IEEE.

Eurostat. (2019). Dane z Eurostat.

Foryś, I. (2010). Wykorzystanie metod taksonomicznych do wyboru obiektów podobnych w procesie wyceny lokali mieszkalnych. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, *18*(1), 95–105.

Jaworowska, M. (2012). PLC standardem przyszłości. 22.03.2013.

Kaliski, M. & Frączek, P. (2012). Rozwój energetyki jądrowej a bezpieczeństwo energetyczne. *Rynek Energii*, *2*, 15–23.

Kaliski, M., Frączek, P. & Szurlej, A. (2011). Brytyjskie doświadczenia a zmiana struktury źródeł energii w Polsce. *Polityka energetyczna*, *14*, 141–153.

Krzysko, M., Wołyński, W., Górecki, T. & Skorzybut, M. (2008). *Systemy uczące się–rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*. Wydawnictwa Naukowo-Techniczne Warszawa.

Kubiak, Z. & Urbaniak, A. (2009). Systemy monitorowania zużycia mediów w budynkach. *Rynek Energii*, *85*(5), 22–31.

Larose, D. T. & Wilbik, A. (2013). *Odkrywanie wiedzy z danych: wprowadzenie do eksploracji danych*. Wydawnictwo Naukowe PWN.

Malko, J. (2006). Energetyczna Strategia Unii Europejskiej. *Wokół Energetyki*, (3).

Michalski, D. (2004). Regulacja rynku energii w UE. *Wspólnoty Europejskie*, (1 (147)).

Michalski, D. (2005). Perspektywy tworzenia wspólnego rynku energii elektrycznej w Unii Europejskiej. *Wspólnoty Europejskie*, (1 (158)), 37–49.

Michalski, D. (2013). Rozwój inteligentnych sieci elektroenergetycznych w Unii Europejskiej. *Unia Europejska. pl*, *1*, 40–50.

Migdał-Najman, K. & Najman, K. (2013). Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej. *Zarządzanie i Finanse*, *3*(2), 179–194.

Niedziółka, D. (2010). *Rynek energii w Polsce*. Difin.

Niedziółka, D. (2018). *Funkcjonowanie polskiego rynku energii*. Difin.

Nowak-Brzezińska, A. (2015). Analiza skupień.

Pach-Gurgul, A. (2012). *Jednolity rynek energii elektrycznej w Unii Europejskiej w kontekście bezpieczeństwa energetycznego Polski*. Difin.

Paska, J. & Surma, T. (2013). Polityka energetyczna Polski na tle polityki energetycznej Unii Europejskiej. *Polityka Energetyczna*, *16*.

Sobierajski, M. & Wilkosz, K. (2000). Sieci elektroenergetyczne a rynki energii elektrycznej. Problemy i perspektywy. *Prace Naukowe Instytutu Energoelektryki Politechniki Wrocławskiej. Konferencje*, *91*(34, t. 1).

Swora, M. & Muras, Z. (2010). Prawo energetyczne. *Komentarz, Warszawa*.

Szablewski, A. T. (2012). Liberalizacja a bezpieczeństwo dostaw energii elektrycznej. *Wydawnictwo Key Text, Warszawa*.

Tarnawski, M. & Młynarski, T. (2016). *Źródła energii i ich znaczenie dla bezpieczeństwa energetycznego w XXI wieku*. Difin SA.

Walesiak, M. (2005). Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów.

Walesiak, M. & Dudek, A. (2007). Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych–charakterystyka problemu.

Walesiak, M. & Gatnar, E. (2009). *Statystyczna analiza danych z wykorzystaniem programu R*. Wydawnictwo Naukowe PWN.

Wojtkowska-Łodej, G. (2014). Wyzwania klimatyczne i energetyczne a polityka Unii Europejskiej. *Polityka Energetyczna*, *17*.

Zajkowski, K. & Borowska, A. (2016). Sieci elektroenergetyczne przyszłości oparte na technologii Smart Grid. *Autobusy: technika, eksploatacja, systemy transportowe*, *17*.

Spis tabel

Spis rysunków

Dodatek A

Spis Programów

A.1 Skrypty użyte do przetwarzania danych

```
install.packages("dplyr")
install.packages("tidyr")
install.packages("ggplot2")
install.packages("stats")
install.packages("tidyverse")
install.packages("lubridate")
install.packages("cluster")
install.packages("clusterSim")
install.packages("fpc")
install.packages("mclust")
install.packages("ggplot2")
install.packages("ggdendro")
install.packages("graphics")
install.packages("corrplot")
library(corrplot)
library(graphics)
library(ggdendro)
library(ggplot2)
library(mclust)
library(tidyverse)
library(lubridate)
library(dplyr)
library(tidyr)
library(ggplot2)
library(stats)
library(cluster)
library(clusterSim)
library(fpc)
#Tworzenie zmiennych
#Zmienna X1
zm_1 = energia %>%
  group_by(LCLid) %>% summarise(pkt_1 = mean(KWH.hh..per.half.hour.))
#Zmienna X2
zm_2 = energia %>%
  group_by(LCLid) %>% summarise(pkt_2 = max(KWH.hh..per.half.hour.))
#Zmienna X3
zm_3 = energia %>%
  group_by(LCLid) %>% summarise(pkt_3 = sum(KWH.hh..per.half.hour.))
#Zmienna X4
gd_4 = energia %>%
```

```
  group_by(LCLid, Date) %>% summarise(pkt = max(KWH.hh..per.half.hour.))
zm_4 = gd_4 %>%
  group_by(LCLid) %>% summarise(pkt_4 = mean(pkt))
#Zmienna X5
gd_5 = energia %>%
  group_by(LCLid, Date) %>% summarise(pkt = mean(KWH.hh..per.half.hour.),
      pkt2 = max(KWH.hh..per.half.hour.))
zm_5 = gd_5 %>%
  group_by(LCLid) %>% summarise(pkt_5 = mean(pkt)/max(pkt2))
#Zmienna X6
zm_6 = energia %>%
  group_by(LCLid) %>% summarise(pkt_6 = var(KWH.hh..per.half.hour.))
#Zmienna X7
zm_7 = energia %>%
  group_by(LCLid) %>% summarise(pkt_7 = sd(KWH.hh..per.half.hour.))
#Zmienna X8
gd_8 = energia %>%
  group_by(LCLid, Date) %>% summarise(pkt = max(KWH.hh..per.half.hour.),
  pktt = min(KWH.hh..per.half.hour.))
zm_8 = gd_8 %>%
  group_by(LCLid) %>% summarise(pkt_8 = max(pkt)-min(pktt))
#Zmienna X9
zm_9 = energia %>%
  group_by(LCLid) %>% summarise(pkt_9 = IQR(KWH.hh..per.half.hour.))
#Zmienna X10
zm_10 = energia %>%
  filter(Time == c("10:00:00.0000000","10:30:00.0000000",
                   "11:00:00.0000000","11:30:00.0000000",
                   "12:00:00.0000000")) %>%
  group_by(LCLid) %>%
  summarise(pkt_10 = max(KWH.hh..per.half.hour.)-mean(KWH.hh..per.half.hour
      .))
#Zmienna X11
zm_11 = energia %>%
  filter(Time == c("10:00:00.0000000","10:30:00.0000000",
                   "11:00:00.0000000","11:30:00.0000000",
                   "12:00:00.0000000")) %>%
  group_by(LCLid) %>%
  summarise(pkt_11 = max(KWH.hh..per.half.hour.)-min(KWH.hh..per.half.hour
      .))

#Korelacyjny wykres rozrzutu
pairs(na.omit(analiza), main = "Korelacyjny wykres rozrzutu")

#Korelacje Spearmana
n = cor(na.omit(analiza), method = "spearman")
corrplot.mixed(n, lower.col = "black", number.cex = .7)

#Standaryzacja zmiennych
analiza.stand = scale(na.omit(analiza))

#Wybór zmiennych do klasyfikacji i~wykres osypiska
r1 = HINoV.Mod(analiza.stand, type = "metric", s = 2, 3, distance = "d4",
               method = "pam", Index = "cRAND")
options(OutDec = ",")
plot(r1$stopri[,2], type = "b", pch = 0, xlab = "Numer zmiennej", ylab = "
    topri", xaxt = "n")
axis(1, at = c(1:max(r1$stopri[,1])), labels = r1$stopri[,1])

# Indeks ńCaliskiego-Harbiasza
ch<- pam(analiza.stand_2,5489)
index.G1(analiza.stand_3,ch$clustering)

#Dendrogram
dd <- dist(analiza.stand_2, method = "euclidean")
```

```
hc <- hclust(dd, method = "ward")
plot(hc, main = "Dendrogram", hang = -1)

#Analiza ńskupie
kmeans_fit_2 = kmeans(analiza.stand_2, centers = 5)
kmeans_fit_2$centers
kmeans_fit_2$cluster
kmeans_fit_2$totss
kmeans_fit_2$withinss
kmeans_fit_2$iter
kmeans_fit_2$size
clusplot(analiza.stand_2, kmeans_fit_2$cluster, main = "Analiza ńskupie",
    color = TRUE, shade = TRUE, plotchar = TRUE, lines = 0)

#Wspóczynnik Randa
w_1 = replication.Mod(analiza.stand, v="m", u=2, centrotypes = "centroids",
     distance = "d2", method = "kmeans")
print(w_1$cRand)

#Opis klas
desc = cluster.Description(analiza, kmeans_fit_2$cluster)
```

Program A.1. Kody w R