Witold Roman

Zastosowanie hierarchicznej metody aglomeracyjnej do grupowania państw OECD ze względu na efektywność wykorzystania energii

1. Wstęp

W 2013 r. w zamieszkałych przez 17,7% ludności naszego globu państwach należących do Organizacji Współpracy Gospodarczej i Rozwoju (Organisation for Economic Co-operation and Development – OECD) wytworzono 71,9% światowego PKB[1] oraz dostarczono 39,1% energii. Przyjmując uproszczony wzór na energochłonność PKB[2], Międzynarodowa Agencja Energetyczna podała w *Key World Energy Statistics 2015*[3], że do uzyskania jednego dolara w państwach OECD zużywano przeciętnie 0,13 kg oleju umownego, a na świecie (łącznie z OECD) – 0,24 kg[4]. Działająca od 30 września 1961 r. OECD skupia obecnie 34 wiodące, najlepiej rozwinięte gospodarczo państwa świata, akceptujące zasady demokracji i gospodarki rynkowej. Każde z państw aspirujących do członkostwa musi więc spełniać te warunki; Polska została przyjęta do OECD dopiero 22 listopada 1996 r.

Pomimo rygorystycznych warunków, jakie muszą spełniać państwa członkowskie, OECD nie stanowi jednolitej grupy. Dzieje się tak, chociaż wartości wskaźników ekonomicznych dotyczące członków OECD zwykle są lepsze niż te same wskaźniki dla państw spoza organizacji. Naturalną rzeczą są różnice dotyczące powierzchni, liczby ludności, warunków klimatycznych czy położenia

---

[1] Do obliczenia wartości wskaźnika wykorzystano dolary amerykańskie według kursu z 2005 r. Według kursu PPP (*purchasing power parity*) wartość wskaźnika jest mniejsza (46,7%).

[2] Iloraz dostarczonej energii pierwotnej i wytworzonego PKB.

[3] *Key World Energy Statistics 2015*, International Energy Agency, Paris 2015, s. 49.

[4] Przy zastosowaniu kurs PPP wartości te były bardziej zbliżone do siebie i wyniosły odpowiednio 0,13 i 0,16 kg. Wynika to oczywiście z wyższych wartości PKB gorzej rozwiniętych państw obliczanego według parytetu siły nabywczej, a nie według jednego sztywnego kursu dla wszystkich państw.

geograficznego. Z ekonomicznego punktu widzenia bardziej interesujące są jednak inne różnice, dotyczące np. struktury gospodarki, osiąganej efektywności ekonomicznej czy wielkości PKB przypadającej na jednego mieszkańca.

Analizy ekonomiczne przeprowadzane na podstawie sformalizowanych modeli wymagają odpowiednio przygotowanych danych statystycznych. Do wstępnej obróbki danych można zaliczyć analizę skupień, w wyniku której otrzymuje się podział wyjściowego zbioru obserwacji na grupy (przeważnie rozłączne)[5], takie że obserwacje wewnątrz jednej grupy są według przyjętego kryterium podobne do siebie, a pomiędzy grupami znacznie się różnią. Uzyskane grupy (klastry, skupienia) mogą następnie być wykorzystane jako dane wyjściowe do dalszych analiz. Wnioskowanie, a w szczególności ekstrapolowanie wartości zmiennych przy wykorzystaniu modeli, których parametry zostały oszacowane na podstawie niejednorodnych obserwacji, niesie ze sobą ryzyko błędów. Ze znaczniejszą niejednorodnością obserwacji mogą bowiem wiązać się różnice mechanizmów kształtujących badane zjawisko. Przeprowadzenie grupowania obserwacji na wstępnym etapie modelowania zjawiska ekonomicznego może więc pomóc uniknąć niebezpieczeństwa wyciągnięcia mało precyzyjnych, a nawet błędnych wniosków.

2. Charakterystyka zastosowanych metod analizy skupień

Podstawowy podział metod analizy skupień obejmuje metody niehierarchiczne i hierarchiczne. Wynikiem działania metod niehierarchicznych jest płaski podział zbioru na klastry, z którego nie można bezpośrednio odczytać zależności występujących pomiędzy wyodrębnionymi grupami. Inaczej jest w przypadku metod hierarchicznych, które dzieli się na metody aglomeracyjne i deglomeracyjne[6]. Tutaj stosowane procedury prowadzą do wyszczególnienia podgrup pozostających względem siebie w określonych relacjach, które łatwo odczytać, korzystając z wygodnego sposobu wizualizacji w postaci dendrogramu. Dendrogram to binarna struktura drzewiasta, której liście odpowiadają analizowanym

[5] Odstępstwem jest rozmyta analiza skupień.

[6] K. Migdał-Najman i K. Najman podają obszerną literaturę dotyczącą systematyki metod grupowania. K. Migdał-Najman, K. Najman, *Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej*, http://zif.wzr.pl/pim/2013_3_2_13.pdf [odczyt 19.02.2016].

elementom zbioru wyjściowego, a węzły – grupom; lokalizacja węzłów wskazuje na poziom podobieństwa/niepodobieństwa grup.

Do pogrupowania krajów członkowskich OECD została wykorzystana aglomeracyjna metoda grupowania hierarchicznego. Metoda ta polega na łączeniu w kolejnych iteracjach najbardziej do siebie podobnych podzbiorów w coraz to większe zbiory; na samym początku podzbiorami są wszystkie jednoelementowe składniki zbioru wyjściowego, na końcu zaś uzyskuje się jeden zbiór, złożony ze wszystkich elementów wyjściowych[7].

Zarówno podział na n grup jednoelementowych, jak i jedna grupa n-elementowa nie mówią nic o strukturze analizowanego zbioru. Dlatego też należy ustalić odpowiednią liczbę klastrów. Przy odwołaniu się do wizualizacji przy pomocy dendrogramu oznacza to przycięcie drzewa na pewnym poziomie zależnym od przyjętej liczby grup. Jej ustalenie ma kluczowe znaczenie na tym etapie analizy. W przypadku niezbyt licznego zbioru procedurę grupowania można powtórzyć dla różnej liczby klastrów, w każdym przypadku oceniając jakość podziału, i jako docelowy wybrać ten, który jest oceniony najwyżej.

Do oceny jakości podziału opracowano szereg wskaźników[8], spośród których do celów grupowania państw OECD został przyjęty wskaźnik sylwetki (silhouette) określony wzorem:

$$s(x) = \frac{b(x_i) - a(x_i)}{max\left(a(x_i), b(x_i)\right)}, \tag{1}$$

gdzie:
$a(x_i)$ – średnia odległość elementu i od pozostałych elementów grupy,
$b(x_i)$ – średnia odległość elementu i od elementów najbliższej sąsiedniej grupy.
Obliczenie dla każdego elementu zbioru wartości tego wskaźnika kwantyfikuje trafność jego przypisania do danej grupy. Zakresem wartości, jakie może przyjmować wskaźnik sylwetki, jest przedział (–1; +1), przy czym wartości bliskie +1 świadczą o poprawnym przypisaniu, natomiast –1 – o ewidentnie złym; elementy mające wartość sylwetki bliskie 0 znajdują się na pograniczu grup[9].

---

[7] Podobną intuicję stosuje się w deglomeracyjnej metodzie hierarchicznej. Tutaj, wychodząc z jednej grupy zawierającej wszystkie elementy zbioru, w kolejnych iteracjach dokonuje się podziału na podgrupy, stosując kryterium maksymalnego niepodobieństwa. Na końcu otrzymuje się n podgrup jednoelementowych.

[8] Zob. K. Migdał-Najman, *Ocena jakości wyników grupowania – przegląd bibliografii*, „Przegląd Statystyczny", t. 58, z. 3–4, Polska Akademia Nauk, Warszawa 2011.

[9] Dokładne omówienie wskaźnika sylwetki jest zawarte w: L. Kauffman, P. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*, John Wiley & Sons Inc., New Jersey 2005.

Do syntetycznej oceny grupowania można posłużyć się średnią arytmetyczną indywidualnych wartości wskaźnika obliczonych dla każdego elementu osobno – im jest ona wyższa, tym bardziej poprawne jest grupowanie. Należy jeszcze zwrócić uwagę na liczbę elementów o ewentualnie ujemnej wartości wskaźnika. W pojedynczych przypadkach można poprawić jakość grupowania, przesuwając te elementy do sąsiednich klastrów, jednak większa ich liczba stawia pod znakiem zapytania zasadność uznania wyników analizy za poprawne.

Do przeprowadzenia analizy skupień niezbędne jest przyjęcie miary odległości oraz metody łączenia zbiorów. Do każdego z tych pojęć można podejść na kilka sposobów. Na potrzeby niniejszej analizy zastosowano wariantowo miarę euklidesową (*Euclidean*) oraz miejską (*Manhattan*)[10].

Odległość euklidesową oblicza się ze wzoru:

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

natomiast miejską ze wzoru:

$$d\left(x,y\right)=\sum_{i=1}^{n}\left|x_i-y_i\right|, \tag{3}$$

gdzie:
$d(x, y)$ – odległość między wektorami x i y,
x_i, y_i – wartość i-tej składowej odpowiednio wektora x oraz y.

Nawiązując do metod obliczania odległości pomiędzy wielowymiarowymi elementami zbioru, należy zwrócić uwagę na kwestię porównywalności wymiarów. W praktyce jeden z wymiarów może zdominować pozostałe, wypaczając całkowicie wyniki obliczeń odległości. Do rozwiązania tej kwestii proponuje się standaryzację lub normalizację zmiennych. W naszej dalszej analizie ujednolicenie wpływu zmiennych zostanie ograniczone do ich standaryzacji przeprowadzonej według wzoru:

$$\tilde{x}_{ij}=\frac{x_{ij}-\overline{x}_j}{D_j}, \tag{4}$$

gdzie:
x_{ij} – wartość i-tej składowej j-tej zmiennej,
$\overline{x}_j$ – średnia wartość zmiennej j obliczana według wzoru:

[10] Innymi miarami są np. odległość Czebyszewa, Minkowskiego czy Mahalanobisa.

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\,, \tag{5}$$

D_j – średnia wartość odchylenia bezwzględnego zmiennej j obliczana według wzoru:

$$D_j = \frac{1}{n}\sum_{i=1}^{n}\left|x_{ij} - \overline{x}_j\right|. \tag{6}$$

Kluczowe w kontekście aglomeracyjnych metod grupowania jest łączenie podzbiorów. Spośród różnych podejść do kwestii łączenia najczęściej wykorzystuje się metodę najbliższego sąsiada (*single*), najdalszego sąsiada (*complete*), metodę średnich odległości elementów (*average*) oraz metodę Warda[11]. Te właśnie podejścia wykorzystano wariantowo na potrzeby niniejszego grupowania.

Odległością pomiędzy zbiorami liczoną metodą najbliższego sąsiada jest najmniejsza odległość znajdowana dla wszystkich par (x_i, y_j), gdzie x_i jest i-tym elementem należącym do pierwszego zbioru, natomiast y_j – j-tym elementem należącym do drugiego zbioru. W metodzie najdalszego sąsiada jako odległość pomiędzy zbiorami przyjmuje się największą odległość pomiędzy wszystkimi wspomnianymi parami elementów. W przeciwieństwie do tych obu skrajnych podejść metoda średniej odległości odwołuje się do miary określonej średnią arytmetyczną wszystkich odległości pomiędzy parami elementów obu zbiorów. W kolejnych krokach łączy się te pary podzbiorów, które dzieli najmniejsza odległość. Nieco inne podejście stosuje się w metodzie Warda. Tutaj kryterium łączenia jest minimalna wartość wariancji nowotworzonej grupy. Dostępne w danym kroku podzbiory łączy się parami w nowe grupy i dla każdej z nich oblicza się wartość wariancji. Wśród utworzonych grup znajduje się tę, której wariancja jest najmniejsza. Jest to ostateczna grupa utworzona w danym kroku. Procedurę powtarza się aż do uzyskania grupy obejmującej wszystkie elementy wyjściowego zbioru poddanego analizie skupień.

[11] Bardziej skomplikowane i mniej intuicyjne są metody *flexible* oraz jej szczególny przypadek – *weighted*.

3. Specyfikacja zmiennych wykorzystanych w analizie skupień państw OECD

Podziału państw OECD na grupy można dokonać na podstawie podzbiorów zmiennych, których wartości są notowane i gromadzone przez najważniejsze organizacje międzynarodowe. Wyboru samego zestawu zmiennych należy dokonać w powiązaniu z celem przeprowadzanej analizy skupień. Celem artykułu jest podział państw OECD na grupy, które będą homogeniczne pod względem podaży i wykorzystania energii na tle sytuacji ludnościowej. Należy się spodziewać, że uzyskane skupienia nie będą jednolite pod względem efektywności energetycznej, wyrażającej się np. energochłonnością PKB. Będą więc mogły stanowić punkt wyjścia do dalszych analiz modelowych ukierunkowanych na efektywne wykorzystanie energii.

Najważniejszymi zmiennymi przydatnymi w analizie skupień dla tak postawionego celu w sposób naturalny wydają się: liczba ludności w państwie, produkcja i zużycie energii oraz saldo wymiany energii z zagranicą.

Liczba ludności odwzorowuje potencjał siły roboczej państwa, jak również potencjał intelektualny i związane z nim możliwości wdrażania innowacji. Substytucja innych czynników w krótkim i średnim horyzoncie czasowym nie ma tutaj pełnego zastosowania, a zatem bezpieczniej jest podzielić objęte modelowaniem państwa na grupy, biorąc pod uwagę wielkości ich populacji.

Zmienna zużycie energii ma bezpośredni związek z energochłonnością PKB. Zaspokojenie potrzeb energetycznych państwa zależy z kolei od pozyskania energii i jej salda wymiany z zagranicą[12]. Wszystkie te zmienne kwalifikują się więc do uwzględnienia w przeprowadzanej analizie skupień.

4. Dane wejściowe do analizy skupień

Jako źródło danych wejściowych do procedury grupowania przyjęto bazy danych Banku Światowego. Z ich tablic można bezpośrednio odczytać liczbę ludności, wielkość zużycia energii oraz – w zmiennej import[13] – saldo wymiany z zagranicą. Natomiast zmienna pozyskanie energii nie jest dostępna *explicite*,

[12] Dla uproszczenia pomijamy tu saldo zapasów.

[13] Eksport jest przez Bank Światowy przedstawiany jako import ze znakiem ujemnym.

ale jej wielkość wynika ze zużycia i salda wymiany. Odczytane z tablic wartości zmiennych są zawarte w poniższym zestawieniu.

Tabela 1. Wartości zmiennych wykorzystanych w analizie skupień państw OECD według stanu za 2013 r.

Kraj	Ludność	Zużycie energii w *toe*[14]	Wymiana energii z zagranicą w *toe*
Australia	23 125 868	131 353 501	–218 914 028
Austria	8 479 375	33 520 059	21 265 351
Belgia	11 182 817	56 353 543	40 150 271
Chile	17 619 708	38 638 103	24 704 337
Czechy	10 514 272	41 354 651	11 590 038
Dania	5 614 932	17 614 997	661 993
Estonia	1 317 997	5 879 621	484 276
Finlandia	5 438 972	32 414 182	14 891 723
Francja	65 920 302	253 362 479	116 826 546
Grecja	11 027 549	23 778 624	14 375 173
Hiszpania	46 620 045	116 022 216	82 325 853
Holandia	16 804 432	77 424 137	7 854 644
Irlandia	4 598 294	13 433 236	10 886 148
Islandia	323 764	5 463 823	598 562
Izrael	8 059 500	25 298 047	18 627 777
Japonia	127 338 621	452 546 647	425 390 738
Kanada	35 158 304	254 089 335	–185 595 660
Korea	50 219 669	263 295 088	219 599 629
Luksemburg	543 360	4 074 226	3 940 625
Meksyk	122 332 399	184 714 422	–35 355 746
Niemcy	80 645 605	313 336 361	193 258 631
Norwegia	5 079 623	33 317 986	–157 251 004
Nowa Zelandia	4 442 100	18 885 050	2 965 035
Polska	38 040 196	96 938 777	26 772 389
Portugalia	10 457 295	21 955 299	16 537 120
Słowacja	5 413 393	16 940 448	10 370 182
Słowenia	2 059 953	6 568 748	3 080 308
Stany Zjednoczone	316 497 531	2 202 962 117	329 760 465
Szwajcaria	8 089 346	27 042 909	13 939 418

[14] *Toe* – tona oleju ekwiwalentnego równa $41{,}868 \cdot 10^9$ J.

Kraj	Ludność	Zużycie energii w *toe*	Wymiana energii z zagranicą w *toe*
Szwecja	9 600 379	48 460 866	14 265 947
Turcja	74 932 641	117 376 155	86 704 221
Węgry	9 893 082	22 809 971	12 494 881
Wielka Brytania	64 106 779	191 421 075	81 455 497
Włochy	60 233 948	157 134 968	121 774 141

Źródło: dane Banku Światowego.

5. Platforma zastosowanej analizy skupień

Do przeprowadzenia grupowania wykorzystano funkcję `agnes ()`[15], wchodzącą w skład pakietu *cluster*, zaimplementowanego w środowisku R. Do oceny jakości przyporządkowania analizowanych elementów do określonych klastrów wykorzystano funkcję `silhouette ()`, a do graficznej prezentacji hierarchii grup w postaci dendrogramu – funkcję `plot ()`; obie te funkcje również wchodzą w skład pakietu *cluster*.

Funkcję `agnes ()` wywołuje się z kilkoma argumentami zależnymi od potrzeb. Jej kompletna postać z domyślnymi wartościami argumentów jest następująca:

```
agnes (x, diss=inherits (x, "dist"), metric="euclidean",
   stand=FALSE, method="average", par.method,
   keep.diss=n < 100, keep.data =!diss, trace.lev=0).
```

Na potrzeby naszej analizy funkcję wywołano z następującymi argumentami[16]:
`x` – macierz obserwacji; wartości tej macierzy przedstawia tabela 1;
`diss` – wartość logiczna: FALSE, gdy `x` jest macierzą obserwacji (przypadek naszej analizy), TRUE, gdy `x` jest macierzą niepodobieństw;
`metric` – łańcuch określający sposób mierzenia odległości pomiędzy obserwacjami; w naszej analizie wykorzystano wariantowo odległość euklidesową (*Euclidean*) oraz miejską (*Manhattan*) – obie standardowo zaimplementowane w funkcji `agnes ()`;

[15] Nazwa funkcji `agnes ()` to akronim od Agglomerative Nesting.
[16] Pełny opis: http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/agnes.html [odczyt 28.11.2015].

`stand` – wartość logiczna: TRUE, gdy wartości macierzy x mają być zestandaryzowane przed przeprowadzeniem obliczeń grupujących (opcja wykorzystana w przeprowadzonej analizie), FALSE – w przeciwnym przypadku;
`method` – łańcuch określający metodę grupowania; w analizie grupowania zostały wykorzystane wariantowo cztery metody: najbliższego sąsiada (*single*), najdalszego sąsiada (*complete*), metoda średnich odległości elementów (*average*) oraz metoda Warda.

Służący ocenie grupowania wskaźnik sylwetki (*silhouette*) można obliczyć, korzystając z włączonej do pakietu *cluster* funkcji `silhouette ()`. W najprostszym przypadku[17] składnia tej funkcji jest następująca:
`silhouette (x),`
gdzie: `x` – obiekt mogący przyjmować różną postać, utworzony w wyniku analizy skupień przeprowadzonej przez funkcję zaimplementowaną w środowisku R; dla grup jednoelementowych przyjmuje się wartość wskaźnika sylwetki równą 0, o czym trzeba pamiętać, interpretując średnią wartość wskaźnika sylwetki.
Wynikiem funkcji `silhouette ()` jest obiekt klasy *silhoutte*, zestawiający grupowane elementy, numery klastrów, do których zostały one przypisane, numery klastrów sąsiednich oraz wartości indywidualnych wskaźników sylwetki $s(\cdot)$.

Syntetyczne informacje o jakości przeprowadzonego grupowania można uzyskać, wywołując funkcję `summary (x)`, którego głównym argumentem jest obiekt klasy *silhouette*. Z kolei graficzną postać wyników można przedstawić, korzystając z funkcji `plot (x)`, również podając jako główny argument obiekt klasy *silhouette*[18].

6. Zastosowanie schematu hierarchicznego aglomeracyjnego grupowania państw OECD oraz ocena jakości wyników

Przyjmując jako wyjściowe dane zawarte w tabeli 1, grupowanie przeprowadzono według następującego porządku:

1. Przyjmując argument `metric = "euclidean"`, dokonano grupowania, stosując kolejno metody: najbliższego sąsiada, najdalszego sąsiada, metodę średnich odległości elementów oraz metodę Warda.

[17] Dokładne omówienie funkcji `silhouette ()`: http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/silhouette.html [odczyt 28.11.2015].

[18] W artykule nie wykorzystano możliwości graficznej prezentacji jakości grupowania.

2. Przyjmując argument `metric = "manhattan"`, dokonano grupowania, stosując kolejno metody: najbliższego sąsiada, najdalszego sąsiada, metodę średnich odległości elementów oraz metodę Warda.
3. Dla każdego z 8 grupowań otrzymanych w punktach 1–2 obliczono średnie wartości wskaźnika sylwetki, przyjmując kolejno podział zbioru wyjściowego na 4, 5, 6, 7, 8, 9, 10, 11, 12 i 13 grup. Odwołując się do drzewiastej wizualizacji, składy grup można odczytać po przycięciu dendrogramu na odpowiednim dla danej liczności klastrów poziomie[19]. Średnie wartości wskaźnika przedstawiono w tabelach 2 i 3.
4. Z tabel 2 i 3 odczytujemy, że średnie wartości wskaźnika sylwetki okazały się najwyższe dla podziałów przeprowadzonych metodą Warda; przy zastosowaniu miary euklidesowej wskaźnik osiągnął wartość 0,6170 przy podziale na 5 grup, natomiast przy wyborze miary miejskiej – 0,6329 (również przy podziale na 5 grup).

Tabela 2. Wartości wskaźnika sylwetki – miara euklidesowa

Liczba grup	Metoda			
	single	*complete*	*average*	Ward
4	0,2829	0,5409	0,5538	0,6010
5	0,3841	0,4891	0,3807	0,6170
6	0,4840	0,5893	0,4840	0,5893
7	0,4403	0,6023	0,5828	0,6023
8	0,3924	0,5872	0,5718	0,5872
9	0,3858	0,4307	0,4807	0,4307
10	0,5422	0,4049	0,4449	0,2871
11	0,4148	0,3741	0,4383	0,2613
12	0,3502	0,3675	0,3502	0,2305
13	0,3614	0,3371	0,3371	0,2239

Źródło: obliczenia własne.

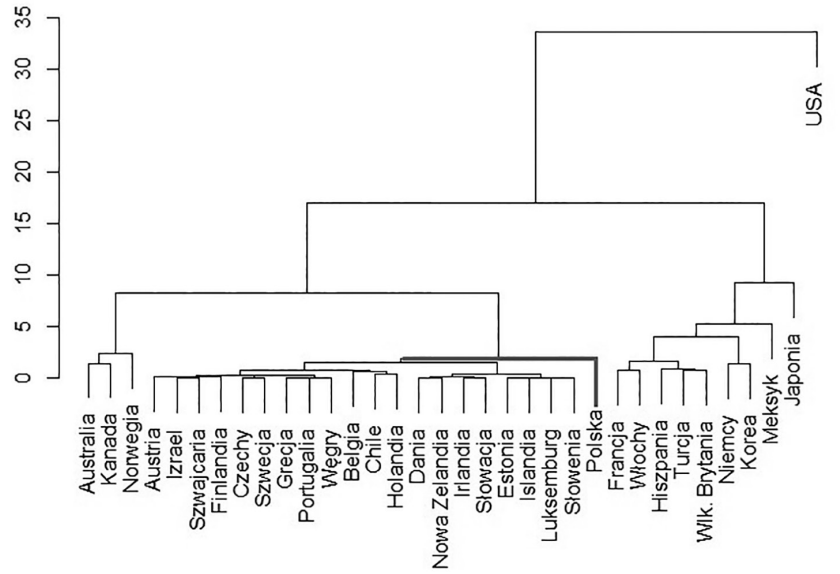
Tabela 3. Wartości wskaźnika sylwetki – miara miejska

Liczba grup	Metoda			
	single	*complete*	*average*	Ward
4	0,3170	0,5655	0,5762	0,5762
5	0,3964	0,6151	0,5348	0,6329

[19] W pakiecie *cluster* przycięcie drzewa realizuje funkcja `cutree ()`.

Liczba grup	Metoda			
	single	*complete*	*average*	Ward
6	0,2285	0,6154	0,6154	0,6154
7	0,3664	0,6232	0,6232	0,6232
8	0,3532	0,5803	0,5803	0,5803
9	0,3070	0,4556	0,5671	0,4556
10	0,5306	0,4413	0,5306	0,4413
11	0,4068	0,4281	0,4068	0,3234
12	0,4010	0,4010	0,4010	0,3102
13	0,3607	0,2831	0,3607	0,2831

Źródło: obliczenia własne.



Rysunek 1. Dendrogram hierarchicznego podziału państw OECD ze względu na efektywność wykorzystania energii metodą Warda przy zastosowaniu miary miejskiej

Źródło: opracowanie własne.

5. Odwołując się do powyższych uwag, dokładniejszej analizie skupień poddano podział na 5 grup dokonany metodą Warda przy zastosowaniu miejskiej miary odległości. Podział ten należy uznać za najlepszy spośród wszystkich przeprowadzonych w niniejszej analizie.

Najważniejsze syntetyczne wyniki najlepszego podziału zostały przedstawione w tabeli 4. Wszystkie 34 państwa OECD zostały podzielone na 5 grup o licznościach: 3, 21, 8, 1, 1, przy czym skupienia jednoelementowe stanowią: Japonia (grupa 4) i Stany Zjednoczone (grupa 5) – państwa o największym w OECD zużyciu energii oraz wyróżniającej się liczbie ludności.

Trzech (z czterech) eksporterów energii netto znalazło się w jednej grupie (1), z tym że jeden z nich (Norwegia) ze względu na niewielką, ujemną wartość wskaźnika sylwetki powinien być raczej przeniesiony do grupy sąsiedniej (2); Norwegia bardzo dużo eksportuje energii, ale – w porównaniu z pozostałymi członkami grupy – mało jej zużywa oraz ma znacznie mniejszą liczbę ludności.

Tabela 4. Podział na grupy z wykorzystaniem algorytmu AGNES i oceny wskaźnika sylwetki

Kraj	Numer grupy	Numer grupy sąsiedniej	Wskaźnik sylwetki	Kraj	Numer grupy	Numer grupy sąsiedniej	Wskaźnik sylwetki
1. Australia	1	2	0,5039	18. Luksemburg	2	1	0,8632
2. Kanada	1	3	0,4441	19. Estonia	2	1	0,8609
3. Norwegia	1	2	–0,1483	20. Islandia	2	1	0,8581
4. Węgry	2	1	0,9072	21. Chile	2	3	0,8504
5. Szwajcaria	2	1	0,9064	22. Belgia	2	3	0,7632
6. Grecja	2	1	0,9054	23. Holandia	2	3	0,7528
7. Portugalia	2	1	0,9042	**24. Polska**	**2**	**3**	**0,4928**
8. Izrael	2	1	0,9031	25. Francja	3	2	0,5790
9. Słowacja	2	1	0,9025	26. Korea	3	4	0,4983
10. Finlandia	2	1	0,8990	27. Wielka Brytania	3	2	0,4876
11. Irlandia	2	1	0,8973	28. Włochy	3	2	0,4563
12. Nowa Zelandia	2	1	0,8906	29. Niemcy	3	4	0,3865
13. Austria	2	1	0,8895	30. Turcja	3	2	0,2163
14. Dania	2	1	0,8851	31. Meksyk	3	2	0,0874
15. Czechy	2	1	0,8832	32. Hiszpania	3	2	0,0526
16. Szwecja	2	1	0,8703	33. Japonia	4	3	0,0000
17. Słowenia	2	1	0,8703	34. Stany Zjednoczone	5	4	0,0000

Źródło: obliczenia własne.

Grupę 3 stanowi osiem państw o znacznej liczbie mieszkańców, o dużym potencjale ekonomicznym i dużym zużyciu energii. Nieco problematyczne jest

przypisanie do tej grupy Meksyku i Hiszpanii; oba państwa mają małe (chociaż dodatnie) wartości wskaźnika sylwetki. Najliczniejsza, licząca 21 elementów, jest grupa 2, do której należy Polska. W jej skład wchodzą państwa rozwinięte i wysokorozwinięte gospodarczo, ale poza Polską nie są to państwa o dużej liczbie ludności i w skali globalnej nie dysponują dużym potencjałem ekonomicznym.

7. Podsumowanie

Przeprowadzona analiza skupień państw należących do OECD z wykorzystaniem aglomeracyjnego hierarchicznego algorytmu (zaimplementowanego w środowisku R w funkcji *agnes* pakietu *cluster*) dała zadowalające wyniki. Dotyczy to zwłaszcza grupy 2, w której znalazła się Polska i w której indywidualne wartości wskaźnika sylwetki kształtowały się w przedziale ⟨0,4928; 0,9072⟩. Jest to grupa liczna, cechująca się dużą jednolitością.

Uzyskanie znacznych rozmiarów grupy, do której trafiła Polska, należy przyjąć z zadowoleniem. Biorąc pod uwagę cel analizy, czyli wyodrębnienie względnie jednolitych państw, które będzie można poddać dalszej analizie modelowej pod względem efektywności energetycznej, należy stwierdzić, że oznacza to otrzymanie bogatszego materiału porównawczego i możliwość uzyskania bardziej wartościowych wyników.

Bibliografia

Biecek P., *Przewodnik po pakiecie R*, Oficyna Wydawnicza GiS, Wrocław 2008.

Energy Efficiency Indicators: Fundamentals on Statistics, International Energy Agency, Paris 2014.

Kauffman L., Rousseeuw P., *Finding Groups in Data. An Introduction to Cluster Analysis*, John Wiley & Sons Inc., New Jersey 2005.

Key World Energy Statistics 2015, International Energy Agency, Paris 2015.

Larose D.T., *Odkrywanie wiedzy w danych*, Wydawnictwo Naukowe PWN, Warszawa 2013.

Migdał-Najman K., *Ocena jakości wyników grupowania – przegląd bibliografii*, „Przegląd Statystyczny", t. 58, z. 3–4, Polska Akademia Nauk, Warszawa 2011.

Źródła sieciowe

http://data.worldbank.org/indicator [odczyt 22.11.2015].

http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/agnes.html [odczyt 28.11.2015].

http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/silhouette.html [odczyt 28.11.2015].

http://zif.wzr.pl/pim/2013_3_2_13.pdf [odczyt 19.02.2016].

* * *

Using the hierarchical agglomerative method to group OECD countries in the context of energy consumption efficiency

Summary

The goal of the paper is clustering OECD states in the context of energy efficiency. For this purpose, the study involved agglomerative hierarchical clustering method of the dataset using the Agglomerative Nesting algorithm implemented as the agnes () function of the *cluster* package running in the R environment. For evaluation of the results the silhouette () function was applied and for the purpose of presentation in the form of dendrogram – the plot.agnes () function; both functions are included in the *cluster* package.

The results of the performed analysis, especially the determined homogeneous group where Poland was placed, can serve as a starting point for further works associated with energy efficiency improvement.

Keywords: cluster, cluster analysis, cluster's silhouette, dendrogram, energy efficiency, energy security, group, hierarchical agglomerative clustering methods, R programming language, R software environment