Witold Roman

Wykorzystanie algorytmu PAM do grupowania najważniejszych gospodarek światowych ze względu na zużycie energii

1. Wstęp

Efektywność zużycia energii jest ważnym czynnikiem wpływającym na stopień zaspokojenia zapotrzebowania społeczeństw na jej nośniki. Wprawdzie światowe zasoby paliw kopalnych w dalszym ciągu są wysokie, a postęp technologiczny pozwala eksploatować coraz trudniej dostępne złoża, jednak nieuchronnie zbliża się moment ich wyczerpania. W rzeczywistości sytuacja może być nawet poważniejsza, bowiem – zgodnie z krzywą Hubberta[1] – wielkość wydobycia nośnika w pewnym momencie osiąga swoje maksimum, którego pomimo zaangażowania coraz lepszych technologii nie udaje się przekroczyć. Przy zwiększającej się liczbie ludności świata oraz dążeniu do rozwoju gospodarczego oznacza to konieczność podejmowania działań zmierzających do znalezienia nowych źródeł energii lub zwiększenia efektywności wykorzystania energii obecnie rozporządzalnej. To drugie działanie wydaje się gwarantować uzyskanie szybszego efektu przy jednoczesnym mniejszym nakładzie środków.

Na szczególną uwagę zasługuje ponadto umieszczenie zasady zrównoważonego rozwoju na liście priorytetów strategii unijnej *Europa 2020*[2], co m.in. oznacza konieczność racjonalnego i oszczędnego korzystania z zasobów oraz zwiększenia efektywności wykorzystania energii. W tym kontekście dużego znaczenia nabiera kwestia optymalizacji struktury zużycia nośników energii. Z polskiej perspektywy obiecującym podejściem do tak opisanego zagadnienia jest przeanalizowanie określonych aspektów naszego systemu energetycznego na tle innych krajów z wykorzystaniem modeli ilościowych, w tym ekonometrycznych. Z tym wiąże się konieczność wyodrębnienia homogenicznych grup krajów, które

[1] http://www.wolfatthedoor.org.uk/mainpages/hubbert.html (data odczytu: 7.07.2014).

[2] http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/priorities/index_pl.htm (data odczytu: 2.08.2014).

są podobne do Polski ze względu na zestaw wybranych czynników istotnych z uwagi na efektywność energetyczną, wnioski wyciągane na podstawie modeli opracowanych dla niejednorodnych krajów mogą być bowiem niepoprawne. Podział najważniejszych gospodarczo krajów świata na grupy z przydzieleniem Polski do jednej z nich jest właśnie głównym celem niniejszego artykułu.

2. Zastosowana metoda i algorytm grupowania

Pojęciem grupowania (analizy skupień, ang. *cluster analysis*) można określić działanie prowadzące do podziału ustrukturyzowanego zbioru danych (obiektów, obserwacji) na podzbiory (zwane grupami, skupieniami) w taki sposób, że elementy wewnątrz wyodrębnionych podzbiorów są podobne do siebie ze względu na określone kryteria, natomiast należące do odrębnych podzbiorów – różnią się. Analiza skupień jest jednym z kilku zadań, które są stawiane przed eksploracją danych (ang. *data mining*), do której rozwoju główny impuls dają obecne możliwości gromadzenia, przechowywania i przetwarzania olbrzymich ilości danych.

Dotychczas opracowano wiele metod grupowania, jednak nawet pobieżne ich omówienie wymagałoby odrębnego obszernego opracowania[3]. Wśród najważniejszych wymienia się k-medoids – niehierarchiczną, iteracyjno-optymalizacyjną metodę podziału, której zastosowanie prowadzi do uzyskania rozłącznych grup. Jest ona rozwinięciem prostszej metody k-średnich (ang. *k-means*)[4] i – w przeciwieństwie do niej – skutecznie radzi sobie np. z problemem obiektów oddalonych (ang. *outliers*).

Metoda k-medoids polega na wyszukaniu wśród analizowanych obiektów takich *k* medoidów (obiektów reprezentujących skupienia), by zminimalizować sumę odległości (niepodobieństw, ang. *dissimilarities*) wszystkich elementów niebędących medoidami od najbliższych im medoidów. Grupę stanowią więc

[3] Obszerny zestaw publikacji dotyczących analizy skupień dołączony jest do artykułu K. Migdał-Najman, *Ocena jakości wyników grupowania – przegląd bibliografii*, „Przegląd Statystyczny", t. 58, z. 3–4, Polska Akademia Nauk, Warszawa 2011.

[4] Metoda k-średnich została opisana np. w: D.T. Larose, *Odkrywanie wiedzy w danych*, Wydawnictwo Naukowe PWN, Warszawa 2013. Została ona upowszechniona w 1967 r. przez J.B. MacQueena (http://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992, data odczytu: 8.07.2014), chociaż jako autora pomysłu podaje się H. Steinhausa (1956) (https://archive.org/stream/arxiv-1201.6082/1201.6082_djvu.txt, data odczytu: 8.07.2014).

medoid i te obiekty, które są od niego w mniejszej odległości niż od pozostałych medoidów. W 1987 r. L. Kaufman i P.J. Rousseeuw opracowali algorytm PAM (*Partitioning Around Medoids*)[5], który jest głównym algorytmem realizującym metodę k-medoids. Jego złożoność obliczeniowa – nawet przy obecnej wydajnej technologii przetwarzania danych – jest oceniana jako wysoka[6], a więc nie może być efektywnie wykorzystywana w przypadku dużych zbiorów. Jednak przy niewielkiej liczbie obserwacji (w naszym przypadku ograniczonej liczbą krajów objętych analizą) algorytm PAM może być z powodzeniem zastosowany bez obawy o czas przetwarzania. Do grupowania większych zbiorów danych w 1990 r. Kaufman i Rousseeuw opracowali algorytm CLARA (*Clustering for Large Applications*)[7], a w 1994 r. R.T. Ng i J. Han – algorytm CLARANS (*Clustering Large Applications based upon Randomized Search*)[8].

Ideę algorytmu PAM można zawrzeć w następujących punktach:

- określiwszy liczbę grup podziału k, ustalić inicjalny zbiór k obiektów-medoidów i w sposób jednoznaczny przypisać do każdego z nich najbardziej do nich podobne (tj. o najmniejszym niepodobieństwie) obiekty niebędące medoidami; wstępny zbiór k medoidów może być ustalony w sposób losowy, ale już w tym etapie lepiej jest wykonać procedurę dającą możliwość trafniejszego wyboru (jest to tzw. faza budowy, ang. *Build Phase*);
- rozważając wszystkie kombinacje par obiektów medoid–niemedoid, ustalić, jak ewentualne przeniesienie niemedoida do zbioru medoidów i *vice versa* wpłynie na jakość grupowania; jeśli jakość poprawi się, dokonać zamiany i ponownie rozważyć kombinacje par, uwzględniając aktualny zbiór medoidów; procedurę powtarza się do osiągnięcia stabilizacji (jest to tzw. faza zamiany, ang. *Swap Phase*).

Po dokonaniu grupowania celowe jest ocenienie jego jakości. Można tu posłużyć się tzw. sylwetką (ang. *silhouette*) $s(x_i)$ obliczaną dla każdego obiektu x_i. Najpierw dla x_i znajduje się jego średnią odległość $a(x_i)$ od pozostałych obiektów grupy, do której został przydzielony, a następnie wybiera się minimalną wartość $b(x_i)$ spośród obliczonych odległości od x_i do każdej spośród pozostałych grup

---

[5] Szczegółowe omówienie PAM można znaleźć np. w: L. Kauffman, P. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*, John Wiley & Sons Inc., New Jersey 2005.

[6] Dla jednej iteracji jest to $O(k\ (n-k)^2)$, gdzie n – liczba obiektów, k – przyjęta *a priori* liczba grup. W przypadku metody k-średnich złożoność wynosi $O\ (kn)$.

[7] Złożoność algorytmu CLARA wynosi $O(ks^2 + k\ (n-k))$, gdzie n – liczba obiektów, k – przyjęta *a priori* liczba grup, s – rozmiar próbek.

[8] Złożoność algorytmu CLARANS to ok. $O\ (n^2)$.

osobno; odległość x_i od danej grupy oblicza się jako średnią odległość od x_i do wszystkich elementów tej grupy. Obie wielkości zestawia się we wzorze:

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

otrzymując wartość sylwetki dla danego obiektu x_i. Ma ona prostą interpretację: obiekty, dla których wskaźnik jest bliski 1, zostały trafnie zgrupowane, pozostałe (o wartości ok. 0 i ujemnej) prawdopodobnie trafiły do złych grup.

Analizę sylwetek podziału można wykorzystać do heurystycznego ustalenia liczby skupień: zmieniając k (np. począwszy od 2), oblicza się wartości wskaźnika sylwetek i wybiera takie k, dla którego średnia szerokość sylwetki s (ang. *average silhouette width*) jest największa; przez średnią szerokość sylwetki rozumie się średnią wartość sylwetki obiektów skupionych w danej grupie.

3. Zastosowana platforma przetwarzania danych

Na potrzeby niniejszego artykułu do przeprowadzenia analizy skupień i pogrupowania krajów został wykorzystany algorytm PAM oprogramowany w języku R i uruchomiony na macierzystej platformie R. Algorytm PAM został zaimplementowany w postaci funkcji `pam()`[9], wchodzącej w skład pakietu *cluster*. Pakiet ten należy zainstalować osobno, nie wchodzi on bowiem w skład podstawowych składników instalowanych wraz z systemem R. Przed uruchomieniem funkcji `pam()` pakiet *cluster* musi zostać załadowany do pamięci komputera.

Funkcję `pam()` wywołuje się z kilkoma argumentami, których wartości domyślne przedstawione są poniżej:

```
pam (x, k, diss=inherits (x,"dist"), metric="euclidean",
medoids=NULL, stand=FALSE, cluster.only=FALSE, do.swap=TRUE,
keep.diss=!diss &&!cluster.only && n<100, keep.data=!diss
&&!cluster.only, pamonce=FALSE, trace.lev=0)
```

gdzie:

`x` – macierz danych, która – w zależności od wartości argumentu `diss` – może być macierzą obserwacji lub odmienności; wiersze macierzy `x` odpowiadają

[9] http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/pam.html (data odczytu: 15.06.2014).

obserwowanym obiektom, natomiast kolumny – opisującym je atrybutom (wymagane jest to, aby były to atrybuty numeryczne),
`k` – liczba grup,
`diss` – wartość logiczna `TRUE`, gdy `x` jest macierzą odmienności, wartość logiczna `FALSE` – w przeciwnym przypadku,
`inherits (x, "dist")` – funkcja, która zwraca wartość logiczną `TRUE`, jeśli `x` jest klasy `dist`, oraz wartość logiczną `FALSE` – w przeciwnym przypadku,
`metric` – metryka odległości, która ma być wykorzystana przy obliczaniu odmienności pomiędzy obserwacjami; obecnie standardowo w funkcji `pam()` dostępne są metryki:
euklidesowa (*euclidean*):

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

i miejska (*manhattan*)[10]:

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

gdzie:
$d(x,y)$ – odległość między wektorami x i y,
x_i – i-ta współrzędna wektora x,
y_i – i-ta współrzędna wektora y,
`medoids` – k-elementowy wektor liczb naturalnych określających numery obiektów przyjętych jako medoidy inicjalne w przypadku rezygnacji z *Build Phase* lub `NULL` – w przeciwnym przypadku,
`stand` – wartość logiczna `TRUE`, gdy elementy macierzy `x` zostały zestandaryzowane przed obliczeniem odmienności, wartość logiczna `FALSE` – w przeciwnym przypadku[11]; standaryzację przeprowadza się zgodnie z wzorem:

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

[10] W przypadku metryki miejskiej używa się niekiedy nazwy „metryka taksówkowa".
[11] Wybór odpowiedniej wartości argumentu stand pozwala uniknąć nadmiernego wpływu zmiennych o większym zakresie zmienności na wyniki grupowania.

gdzie:

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

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

`cluster.only` – wartość logiczna `TRUE`, gdy działanie `pam ()` ogranicza się tylko do grupowania, wartość logiczna `FALSE` – gdy mają być obliczone również pozostałe wartości funkcji,
`do.swap` – wartość logiczna `TRUE`, gdy *Swap Phase* ma być przeprowadzona, wartość logiczna `FALSE` – w przeciwnym przypadku,
`keep.diss, keep.data` – wartość logiczna wskazująca, czy obliczone odmienności i/lub dane wejściowe macierzy `x` mają być przechowywane z wynikami,
`pamonce` – wartość logiczna lub jedna z liczb ze zbioru {0; 1; 2}; zadeklarowana wartość decyduje o liczbie kroków *Swap Phase* (`pamonce = FALSE` oznacza pełne wykonanie fazy zamiany).

Wynikiem działania funkcji `pam()` jest obiekt klasy pam. Dobierając odpowiednie wartości argumentów, z jakimi wywołuje się funkcję `pam()`, można sterować jej działaniem oraz zakresem wyświetlanych na wyjściu informacji. Najważniejsza jest informacja o podziale analizowanych obiektów na grupy, podawana w formie n-wymiarowego wektora `clustering`, którego współrzędne reprezentują poszczególne elementy zbioru, przy czym wartościami tych współrzędnych są numery grup, do których kolejne obiekty zostały przypisane. Ważną wartością funkcji `pam()`, służącą do oceny jakości przeprowadzonej analizy skupień, jest lista `silinfo`, która podaje najistotniejsze informacje o sylwetce podziału (m.in. średnią wartość wskaźnika sylwetki dla całego zbioru, dla grup oraz dla pojedynczych obiektów).

Uzupełnieniem zwracanych przez funkcję `pam()` wyników liczbowych, a zarazem poglądową formą prezentacji oceny jakości analizy skupień mogą być wykresy wskaźnika sylwetki grupowania. Do ich sporządzenia można zastosować funkcję `plot ()`, którą w najprostszy sposób wywołuje się z podaniem jednego argumentu – obiektu klasy silhouette – wyniku zwracanego przez funkcję `silhouette()`. Tę z kolei można wywołać, podając jako argument obiekt klasy pam, uzyskany przez funkcję `pam()`. Wynikiem funkcji `silhouette()` jest macierz, której kolumny tworzą kolejno: numery obiektów, numery przyporządkowanych im grup, numery grup sąsiednich oraz odpowiadające im wartości wskaźnika sylwetki. Liczba wierszy jest równoliczna ze zbiorem analizowanych obiektów.

Dodatkowe informacje o dokonanym podziale można uzyskać, wywołując funkcję `summary()` z argumentem klasy pam lub silhouette.

4. Ustalenie listy atrybutów

Zobiektywizowanie doboru atrybutów charakteryzujących analizowany zbiór obiektów jest trudne, jeśli nie niemożliwe do realizacji. Ustalając listę tych atrybutów, należy więc kierować się przede wszystkim szczegółową wiedzą o przedmiocie badania. Uzasadniona merytorycznie wstępna lista może zawierać większą liczbę atrybutów, którą ewentualnie można zmienić/zmniejszyć, o ile uzyskiwane oceny grupowania nie będą zadowalające.

Do pogrupowania gospodarek narodowych pod kątem dalszej analizy efektywności energetycznej wstępnie zostało przyjętych następujących dziewięć atrybutów: powierzchnia kraju, produkcja energii, import energii netto[12], zużycie energii, zużycie energii *per capita*, PKB, PKB *per capita*, liczba osób pracujących oraz liczba ludności.

5. Źródło danych oraz przygotowanie danych do analizy skupień

Spośród danych najbardziej renomowanych ośrodków statystycznych na świecie do analizy przyjęto opublikowane przez Bank Światowy (World Bank)[13] wartości atrybutów odnoszące się do 56 państw. Do listy obiektów włączono ważniejsze gospodarki światowe według klasyfikacji przyjętej w *Key World Energy Statistics 2012* – publikacji Międzynarodowej Agencji Energetycznej (International Energy Agency – IEA). Są to:

– kraje OECD: Australia, Austria, Belgia, Kanada, Chile, Czechy, Dania, Estonia, Finlandia, Francja, Niemcy, Grecja, Węgry, Islandia, Irlandia, Izrael, Włochy, Japonia, Korea, Luksemburg, Meksyk, Holandia, Nowa Zelandia, Norwegia, Polska, Portugalia, Słowacja, Słowenia, Hiszpania, Szwecja, Szwajcaria, Turcja, Wielka Brytania i Stany Zjednoczone;

[12] Wartość wskaźnika „import energii netto" ze znakiem ujemnym oznacza „eksport energii netto" (jest to konwencja przyjęta np. przez Bank Światowy).

[13] http://data.worldbank.org/indicator (data odczytu: 20.07.2014).

– kraje Bliskiego Wschodu (bez Syrii[14]): Bahrajn, Iran, Irak, Jordania, Kuwejt, Liban, Oman, Katar, Arabia Saudyjska, Zjednoczone Emiraty Arabskie i Jemen;
– państwa należące do Unii Europejskiej, ale niebędące członkami OECD: Bułgaria, Cypr, Łotwa, Litwa, Malta, Rumunia;
– inne ważne gospodarki: Brazylia, Chiny, Indie, Indonezja, Rosja.

Biorąc pod uwagę możliwie najbardziej kompletny zbiór wartości atrybutów wyjściowych, do analizy przyjęto dane za 2012 r.[15] Jednak nawet dla tego roku bazy danych Banku Światowego wykazują braki – dla 22 krajów wartości czterech atrybutów (produkcja, import, zużycie oraz zużycie energii *per capita*) musiały być imputowane. W tym celu zastosowano prostą ekstrapolację liniową na podstawie szeregów czasowych z lat 2008–2011.

6. Wyniki analizy skupień

Na bazie wstępnego zestawu atrybutów zostały przeprowadzone grupowania (przy założeniu dziewięciu wartości k: od $k = 2$ do $k = 10$ oraz obu dostępnych dla funkcji `pam()` miar odległości: euklidesowej i miejskiej). Tabela 1 przedstawia dla każdej kombinacji „liczba skupień–miara odległości" po trzy wartości pomocne przy ocenie jakości grupowań. Są to: średnia szerokość sylwetki s, liczność największej grupy w każdym skupieniu oraz liczba obiektów o ujemnej wartości wskaźnika sylwetki $s(\cdot)$.

Skupienie zbyt wielu obiektów w jednej grupie, mała szerokość sylwetki czy duża liczba obiektów o ujemnej wartości $s(\cdot)$ stawiają pod znakiem zapytania przydatność dokonanego grupowania. Rozpatrując wyniki analizy skupień na podstawie wstępnych dziewięciu atrybutów, dostrzega się fakt, że grupowanie należy poprawić. W celu poprawy jakości efektów grupowania zdecydowano się zredukować listę atrybutów. Pomijając cztery z nich (import energii, zużycie energii *per capita*, PKB *per capita* oraz wielkość siły roboczej), których wartości można obliczyć na podstawie wartości pozostałych atrybutów lub które są silnie skorelowane z którymś z pozostałych, uzyskano względnie zadowalające wyniki

[14] Syria została wyłączona z analizowanego zbioru ze względu na nieakceptowalną niekompletność dostępnych danych.

[15] Dane za 2013 r. są opublikowane dla atrybutów: powierzchnia i liczba ludności dla wszystkich 56 krajów oraz PKB i PKB *per capita* dla 47 krajów.

podziału. Wielkości charakteryzujące grupowania przeprowadzone na podstawie pięciu atrybutów (powierzchnia kraju, produkcja energii, zużycie energii, PKB, liczba ludności) zostały zebrane w tabeli 2. Wśród wszystkich dziewięciu podziałów na uwagę zasługuje podział na osiem grup, do którego realizacji wykorzystano euklidesową miarę odległości. Średnia szerokość sylwetki s dla tej grupy wynosi 0,41, największa grupa obejmuje 25 krajów i tylko w jednym przypadku $s(\cdot) < 0$.

Tabela 1. Wybrane wartości charakteryzujące jakość grupowania na podstawie wstępnego zestawu atrybutów

Liczba grup k	Miara					
	euklidesowa			miejska		
	średnia szerokość sylwetki s	liczność największej grupy	liczba obiektów o wartości $s(\cdot) < 0$	średnia szerokość sylwetki s	liczność największej grupy	liczba obiektów o wartości $s(\cdot) < 0$
2	0,75	53	1	0,78	54	0
3	0,37	45	4	0,34	42	11
4	0,37	45	4	0,30	42	4
5	0,42	42	3	0,33	38	5
6	0,42	42	2	0,31	40	3
7	0,18	23	12	0,27	30	6
8	0,18	15	11	0,16	16	6
9	0,21	17	6	0,18	16	5
10	0,22	17	5	0,22	17	3

Źródło: obliczenia własne.

Wyniki podziału analizowanego zbioru krajów na grupy wraz z wartościami wskaźnika sylwetki każdego z nich i wyszczególnionymi medoidami są zawarte w tabeli 3. Ponadto wywołując funkcję `summary()`, uzyskano dodatkowe informacje o jakości przeprowadzonej analizy skupień:

1) średnie szerokości sylwetki w grupach: I – 0,4097; II – 0,6596; III – 0,5608; IV – 0,0000; V – 0,1210; VI, VII, VII – 0,0000;
2) łączna średnia szerokość sylwetki w całym zbiorze – 0,4115.

Problemem do rozstrzygnięcia jest ujemna wartość szerokości sylwetki obliczona dla Korei (pozycja 53). W takiej sytuacji można zaproponować przesunięcie obiektu do grupy sąsiedzkiej – w przypadku Korei do grupy I. Ostatecznie grupa I obejmuje 25 krajów plus dołączona Korea, II – 3 kraje, III – 18, IV– 1 (Chiny),

V – 5, grupy VI, VII i VIII – po jednym kraju (odpowiednio: Indie, Rosja i Stany Zjednoczone). Polska znalazła się w grupie I.

Tabela 2. Wybrane wartości charakteryzujące jakość grupowania na podstawie końcowego zestawu atrybutów

Liczba grup k	Miara					
	euklidesowa			miejska		
	średnia szerokość sylwetki s	liczność największej grupy	liczba obiektów o wartości $s(\cdot) < 0$	średnia szerokość sylwetki s	liczność największej grupy	liczba obiektów o wartości $s(\cdot) < 0$
2	0,82	54	0	0,85	54	0
3	0,25	31	9	0,67	49	1
4	0,34	26	4	0,66	49	1
5	0,33	26	4	0,26	25	4
6	0,35	26	1	0,28	25	2
7	0,40	25	1	0,33	24	4
8	0,41	25	1	0,35	24	4
9	0,32	15	7	0,39	23	2
10	0,36	16	5	0,31	16	8

Źródło: obliczenia własne.

Tabela 3. Podział na grupy z wykorzystaniem algorytmu PAM i oceny wskaźnika sylwetki

Kraj	Numer grupy	Numer grupy sąsiedniej	Wskaźnik sylwetki	Kraj	Numer grupy	Numer grupy sąsiedniej	Wskaźnik sylwetki
1. Szwecja	1	3	0,5996	29. Czechy	3	1	0,7175
2. Rumunia	1	3	0,5849	30. Dania	3	1	0,7138
3. Szwajcaria	1	3	0,5829	31. Cypr	3	1	0,7082
4. Słowacja	1	3	0,5813	32. Estonia	3	1	0,7038
5. Słowenia	1	3	0,5727	**33. Finlandia**	3	1	0,7019
6. ZEA	1	3	0,5560	34. Bułgaria	3	1	0,6959
7. Portugalia	1	3	0,5530	35. Chile	3	1	0,6806
8. Jemen	1	3	0,5436	36. Bahrajn	3	1	0,6616
9. Turcja	1	3	0,5339	37. Belgia	3	5	0,6603
10. Polska	1	3	0,5326	38. Grecja	3	1	0,6593

Kraj	Numer grupy	Numer grupy sąsiedniej	Wskaźnik sylwetki	Kraj	Numer grupy	Numer grupy sąsiedniej	Wskaźnik sylwetki
11. Oman	1	3	0,5207	39. Austria	3	1	0,6509
12. Hiszpania	1	3	0,5168	40. Węgry	3	1	0,6386
13. Katar	1	3	0,4878	41. Islandia	3	1	0,6007
14. Nowa Zelandia	1	3	0,4684	42. Irlandia	3	1	0,3991
15. Norwegia	1	3	0,4494	43. Irak	3	1	0,3450
16. Holandia	1	3	0,4245	44. Izrael	3	1	0,3368
17. Malta	1	3	0,3285	45. Jordania	3	1	0,1156
18. Luksemburg	1	3	0,2854	46. Iran	3	1	0,1050
19. Arabia Saud.	1	5	0,2455	**47. Chiny**	4	8	0,0000
20. Wlk. Brytania	1	5	0,2387	48. Niemcy	5	3	0,3605
21. Litwa	1	3	0,2290	49. Japonia	5	1	0,3091
22. Liban	1	3	0,1592	**50. Francja**	5	3	0,2322
23. Meksyk	1	5	0,1327	51. Włochy	5	3	0,0100
24. Łotwa	1	3	0,0769	52. Indonezja	5	3	0,0002
25. Kuwejt	1	3	0,0393	53. Korea	5	1	–0,1860
26. Kanada	2	5	0,7140	**54. Indie**	6	5	0,0000
27. Australia	2	3	0,6514	**55. Rosja**	7	2	0,0000
28. Brazylia	2	5	0,6099	**56. USA**	8	4	0,0000

Źródło: obliczenia własne.

7. Podsumowanie

Uzyskane wyniki analizy skupień mogą być wykorzystane do badań związanych z zaspokojeniem potrzeb energetycznych Polski. Dalsze analizy można przeprowadzić na podstawie danych porównawczych z innych gospodarek tej samej grupy, w której skład wchodzą kraje podobne do Polski ze względu na wybrane cechy istotne w kontekście problematyki energetycznej (w szczególności efektywności wykorzystania energii) rozpatrywanej w skali makro; cechy te są reprezentowane przez atrybuty, które zostały wykorzystane w algorytmie PAM. W składzie grupy znalazły się kraje o zróżnicowanej energochłonności PKB – od 0,06 toe/tys. USD $_{(2005)}$ dla Szwajcarii i 0,08 toe dla Wielkiej Brytanii po 0,57 toe dla Omanu i 0,40 toe dla Arabi Saudyjskiej; wartość tego wskaźnika dla Polski

to 0,24 toe[16]. Generalnie większą energochłonność wykazują duzi producenci nośników energii oraz kraje o niższym poziomie rozwoju gospodarczego. Są jednak i wyjątki, do których należy np. Wielka Brytania czy Norwegia.

Ukierunkowana na efektywność energetyczną analiza sytuacji Polski na tle pozostałych krajów grupy może doprowadzić do interesujących wniosków dotyczących zmian, których celem byłoby lepsze wykorzystanie nośników energii, a zatem i poprawienie bezpieczeństwa energetycznego naszego kraju.

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 2012, International Energy Agency, Paris 2012.

Key World Energy Statistics 2013, International Energy Agency, Paris 2013.

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 (data odczytu: 20.07.2014).

http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/priorities/index_pl.htm (data odczytu: 2.08.2014).

http://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992 (data odczytu: 8.07.2014).

http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/pam.html (data odczytu: 15.06.2014).

http://www.wolfatthedoor.org.uk/mainpages/hubbert.html (data odczytu: 7.07.2014).

https://archive.org/stream/arxiv-1201.6082/1201.6082_djvu.txt (data odczytu: 8.07.2014).

[16] Dane za 2012 r. wg IEA.

* * *

Using the PAM algorithm to group the most important global economies in the context of energy consumption

Summary: The purpose of the paper is to draft a typology of countries in the context of meeting their energy demand as well as to define the current position of Poland. In order to pursue that purpose, the clustering method was applied; the countries were divided into homogeneous groups and Poland was placed in one of them. The obtained results can serve as a starting point for other elaborations on ensuring energy security, fulfilling EU obligations, optimisation of structure of the energy sources in use, as well as energy efficiency improvement.

Within the clustering method, the Partitioning Around Medoids algorithm was used. PAM was developed by Leonard Kaufman and Peter J. Rousseeuw in 1987 and it was implemented as the `pam()` function of the `cluster` package, which is run in the software environment; R PAM is the most common realisation of non-hierarchical, k--medoids clustering method. For evaluation of clustering, the `silhouette()` function was applied, which is also included in the `cluster` package.

Keywords: cluster analysis, energy security, energy efficiency, data mining, group, R programming language, cluster, medoid, non-hierarchical clustering methods, cluster's silhouette, R software environment