PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Joanna Trzęsiok
Uniwersytet Ekonomiczny w Katowicach

TAKSONOMICZNA ANALIZA KRAJÓW POD WZGLĘDEM DZIETNOŚCI KOBIET ORAZ INNYCH CZYNNIKÓW DEMOGRAFICZNYCH

Streszczenie: W pracy przeprowadzono analizę taksonomiczną 180 państw pod względem wybranych czynników demograficznych. Początkowo w zbiorze danych kraje były charakteryzowane przez 11 zmiennych. Jednak z uwagi na słabą jakość uzyskiwanych podziałów zastosowano procedurę eliminacji cech i do ostatecznego badania wybrano tylko cztery z nich: dzietność kobiet, umieralność dzieci przed ukończeniem 5 lat (na 1000 urodzonych), procent populacji w przedziale wiekowym $\langle 5,20)$ oraz procent populacji w wieku powyżej 60 lat. Pokazano, że zmienne te miały istotny wpływ na poprawę jakości grupowania. W badaniu wykorzystano 4 wybrane metody taksonomiczne. Najlepszą strukturę klas odkryła metoda *k*-medoidów, dzieląc badane państwa na trzy skupiska.

Słowa kluczowe: analiza taksonomiczna, demografia, eliminacja zmiennych.

1. Wstęp

Zgodnie z prognozą opublikowaną przez Central Intelligence Agency (CIA), współczynnik dzietności w Polsce w roku 2013 będzie wynosić 1,31 (urodzonych dzieci przypadających na jedną kobietę) i będzie on niższy od współczynnika dzietności w Chinach (1,55), które propagują politykę jednego dziecka. Ponadto jedynie 6% krajów na świecie będzie miało niższy współczynnik dzietności kobiet niż Polska. Ten problem demograficzny nie jest jedynym, z którym boryka się nasz kraj. Innymi zagadnieniami często podnoszonymi przez demografów są chociażby dylematy starzejącego się społeczeństwa. Problemy te nie są również specyfiką Polski. Nasuwa się więc pytanie, do jakich państw Polska jest podobna pod względem wybranych czynników demograficznych, takich jak wspomniana dzietność kobiet, struktura wiekowa społeczeństwa, wskaźnik wzrostu populacji czy oczekiwana długość życia.

Celem artykułu było przeprowadzenie analizy taksonomicznej, wykorzystującej wielowymiarowe metody statystyczne, oraz sklasyfikowanie krajów pod względem

wybranych zmiennych demograficznych. Do analizy początkowo wybrano 11 zmiennych, by ostatecznie przeprowadzić procedurę eliminacji cech i uwzględnić w badaniu tylko te z nich, które mają istotny wpływ na jakość podziału.

2. Procedura badawcza

Do analizy wykorzystano zbiór danych skonstruowany na podstawie zmiennych i ich realizacji z roku 2012, udostępnionych na portalu www.gapminder.org[1]. Obiektami w badaniu było 180 państw[2]. W początkowym etapie każde państwo scharakteryzowane zostało przez 11 zmiennych demograficznych:

X_1 – współczynnik dzietności kobiet[3],

X_2 – oczekiwaną długość życia,

X_3 – umieralność dzieci do 5. roku życia (na 1000 urodzeń),

X_4 – wskaźnik wzrostu populacji,

X_5 – współczynnik maskulinizacji (liczba mężczyzn przypadających na 100 kobiet),

X_6 – procent populacji zamieszkującej na terenie zurbanizowanym,

w tym cechy przedstawiające strukturę wiekową społeczeństwa danego kraju:

X_7 – procent ludności w przedziale wiekowym $\langle 0,5)$,

X_8 – procent ludności w przedziale wiekowym $\langle 5,20)$,

X_9 – procent ludności w przedziale wiekowym $\langle 20,40)$,

X_{10} – procent ludności w przedziale wiekowym $\langle 40,60)$,

X_{11} – procent ludności w wieku powyżej 60 lat.

Wszystkie zmienne mierzone były na skali ilorazowej, zatem jako formułę normalizacyjną zastosowano standaryzację zerowaną.

Do podziału zbioru państw na skupienia wykorzystano cztery metody taksonomiczne, które zaliczane są do podstawowych metod klasyfikacji [Walesiak, Gatnar (red.) 2009, s. 413]:

---

[1] Portal www.gapminder.org upowszechnia wiedzę z obszaru zdrowia globalnego i rozwoju cywilizacji poprzez łatwy dostęp do danych statystycznych, zaczerpniętych głównie ze statystyk ONZ, oraz narzędzi wspomagających wizualizację tych danych. Współtwórcą tego portalu jest prof. Hans Rosling.

[2] Z uwagi na braki w danych niektóre państwa musiały zostać pominięte. Nie uwzględniono również Kataru oraz Zjednoczonych Emiratów Arabskich, ponieważ były to obserwacje oddalone ze względu na realizację zmiennej X_5. Obserwacje te zaburzały wyniki badań.

[3] Terminologię dotyczącą cech demograficznych przyjęto w oparciu o prace [Okólski 2005 oraz Holzer 2003], choć J. Holzer używa zamiennie pojęć „dzietność" oraz „płodność całkowita".

- dwie metody optymalizujące wstępny podział zbioru obiektów: *k*-średnich [MacQueen 1967] oraz *k*-medoidów [Kauffman, Rousseeuw 1990; Jajuga 1993; Pociecha i in. 1986],
- dwie hierarchiczne metody aglomeracyjne: Warda [1963] oraz kompletnego połączenia [Defays 1977; Walesiak, Gatnar (red.) 2009; Kopczewska i in. 2009].

W przeprowadzonej analizie skupień stosowano wymienione metody, dzieląc badany zbiór obiektów na *k* klas, dla $k = 2,\ldots,5$. Wyniki grupowania oceniono za pomocą indeksu sylwetkowego I_S (*Silhouette Indeks*), przyjmując, zgodnie z zaproponowanymi w pracy [Kauffman, Rousseeuw 1990] wartościami progowymi miernika I_S, że jeśli:

- $I_S > 0,5$, to odkryto poważną strukturę klas,
- $I_S > 0,7$, to mamy do czynienia z silną strukturą klas.

3. Identyfikacja zmiennych istotnie wpływających na jakość grupowania

W pierwszym etapie analizy zbudowano wiele modeli taksonomicznych, przyjmując liczbę klas $k = 2,\ldots,5$ oraz wykorzystując cztery wymienione metody. Niestety, wartości indeksu sylwetkowego, obliczonego dla każdego z tych modeli, były niższe od 0,5, co wskazywało na słabą strukturę klas. Z tego też względu w dalszym kroku analizy zastosowano procedurę eliminacji pojedynczych zmiennych [Guyon i in. (red.) 2006], by z całego zestawu cech wybrać tylko te, które będą miały istotny wpływ na jakość grupowania.

Tabela 1. Algorytm procedury eliminacji pojedynczych cech

Dla każdej z wybranych metod taksonomicznych oraz dla zadanej liczby klas k wykonaj następujące kroki:	
Krok 1.	Wykorzystując pełen zestaw zmiennych, podziel zbiór państw D na k klas. Utwórz pomocniczy zbiór S będący kopią zbioru D.
Krok 2.	Poprzez wyłączenie tymczasowo ze zbioru S kolejno każdej ze zmiennych wygeneruj wiele zmodyfikowanych zbiorów danych na bazie S. Podziel tak zmodyfikowane zbiory państw na k klas.
Krok 3.	Oceń jakość każdego podziału uzyskanego w kroku 2. za pomocą indeksu sylwetkowego.
Krok 4.	Zidentyfikuj ten podział zbioru danych z wyłączoną zmienną, dla którego wartość indeksu sylwetkowego jest największa, a następnie usuń ze zbioru S tę zmienną.
Krok 5.	Powróć do kroku 2. i powtarzaj procedurę, dopóki w S pozostaje więcej niż jedna zmienna.
Krok 6.	Z otrzymanego ciągu modeli taksonomicznych (z malejącą liczbą zmiennych) wybierz ten, dla którego wartość indeksu sylwetkowego jest największa.

Źródło: opracowanie własne.

W procedurze eliminacji pojedynczych zmiennych początkowo do podziału obiektów wykorzystano wszystkie zmienne. W każdym kolejnym kroku usuwano jedną zmienną, według ustalonego *a priori* kryterium i ten zmniejszony zbiór cech posłużył do budowy następnych modeli taksonomicznych. Eliminowane były po kolei te zmienne, które miały najmniejszy wpływ na jakość podziału. Kryterium wyboru zmiennej do usunięcia był maksymalny indeks sylwetkowy. Procedurę powtarzano tak długo, aż w zbiorze pozostała tylko jedna zmienna – ta, która miała największy wpływ na jakość grupowania państw. Kroki algorytmu omówionej procedury przedstawiono w tabeli 1.

4. Wyniki analizy

Eliminację pojedynczych zmiennych przeprowadzono 16 razy – dla każdej z wymienionych metod taksonomicznych oraz liczby klas $k = 2,\ldots,5$. Ze względu na ograniczenia objętości tej pracy szczegółowo przedstawiono etapy omawianej procedury tylko w jednym przypadku – dla metody Warda i liczby klas równej 2 (zob. tab. 2). Z ciągu indeksów sylwetkowych, obliczonych dla modeli taksonomicznych, budowanych (metodą Warda z $k = 2$) dla zbioru państw z malejącą liczbą zmiennych, najlepszy jest ten, który uzyskano w ostatnim kroku procedury, czyli $I_S = 0,760$. Oznacza to, że najlepszą strukturę klas otrzymano, grupując państwa na podstawie tylko jednej zmiennej – tej, która pozostała w modelu w ostatnim etapie, czyli X_{11} reprezentującej procent populacji w wieku powyżej 60 lat.

Tabela 2. Wynik działania procedury eliminacji pojedynczych zmiennych w przypadku grupowania państw metodą Warda na 2 skupiska

Etap	Usunięta zmienna	Wartość I_S	Etap	Usunięta zmienna	Wartość I_S
1	Ø	0,362	7	X_1	0,565
2	X_6	0,414	8	X_4	0,586
3	X_5	0,449	9	X_7	0,617
4	X_9	0,487	10	X_{10}	0,680
5	X_3	0,486	11	X_8	**0,760**
6	X_2	0,521	12	X_{11}	

Źródło: opracowanie własne.

Podsumowanie wyników wszystkich wykonanych eksperymentów przedstawiono w tabeli 3.

Wyniki zamieszczone w tabeli 3, pokazują, że we wszystkich badanych przypadkach najlepsze wartości I_S otrzymywano zawsze w ostatnim kroku procedury eliminacji zmiennych, co oznacza, że grupowanie w każdym przypadku odbywało się na podstawie tylko jednej cechy. Warto jednak zauważyć, że zmienne wykorzystane w badaniu, czyli te, które miały istotny wpływ na jakość podziału zbioru

Tabela 3. Wartości indeksu sylwetkowego dla najlepszych podziałów państw (dla różnych wariantów modeli taksonomicznych) uzyskanych z wykorzystaniem procedury eliminacji pojedynczych zmiennych, jak i te zmienne, które zastosowano do tego podziału

Liczba klas \ Metoda		*k*-średnich	*k*-medoidów	Warda	Kompletnego połączenia
$k = 2$	najlepszy I_S	0,715	0,704	**0,760**	0,747
	dla zmiennej	X_3	X_3	X_{11}	X_1
$k = 3$	najlepszy I_S	0,675	0,670	0,659	0,706
	dla zmiennej	X_8	X_8	X_8	X_3
$k = 4$	najlepszy I_S	0,621	0,623	0,615	0,573
	dla zmiennej	X_8	X_8	X_8	X_{11}
$k = 5$	najlepszy I_S	0,584	0,587	0,606	0,610
	dla zmiennej	X_1	X_{10}	X_7	X_3

Źródło: opracowanie własne.

danych, powtarzają się. Najlepszą strukturę skupisk otrzymywano, gdy zbiór państw charakteryzowany był przez: X_1, X_3, X_8 lub X_{11}. Poszukiwanie skupisk w zbiorze państw opisywanych przez te pojedyncze cechy prowadziło do bardzo dobrych wyników. Zachodzi jednak pytanie, jak dobry podział można uzyskać, wprowadzając do modelu taksonomicznego te 4 zmienne jednocześnie.

W kolejnym kroku analizy (za pomocą różnych metod i dla $k = 2,\ldots,5$) dokonano podziału państw charakteryzowanych przez: dzietność kobiet (X_1), umieralność dzieci do 5. roku życia (X_3), procent ludności w przedziale wiekowym $\langle 5,20)$ (X_8) oraz procent populacji w wieku powyżej 60 lat (X_{11}). Oceny jakości podziału, czyli wartości indeksu sylwetkowego obliczonego dla każdego z badanych modeli taksonomicznych, zaprezentowano w tabeli 4.

Tabela 4. Wartości indeksu sylwetkowego mierzącego jakość podziału państw charakteryzowanych przez zmienne: X_1, X_3, X_8 i X_{11}, dla różnych wariantów modeli taksonomicznych

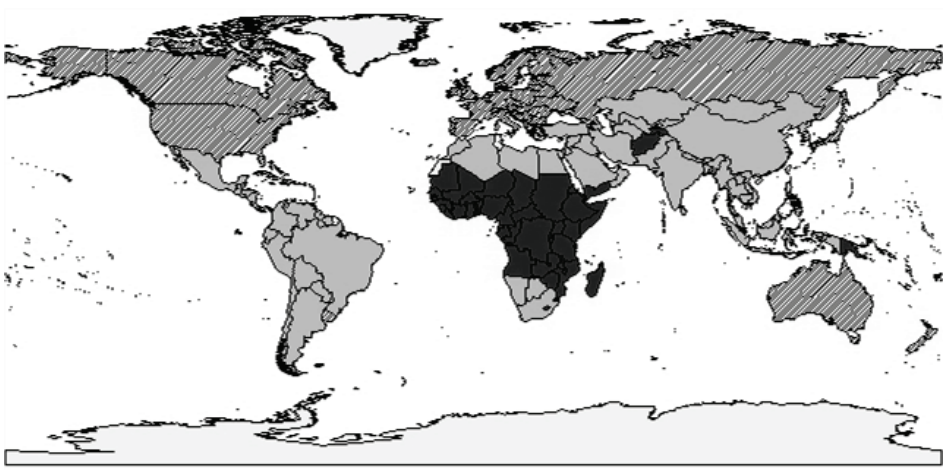
Liczba klas \ Metoda	*k*-średnich	*k*-medoidów	Warda	Kompletnego połączenia
$k = 2$	0,480	0,501	0,500	0,486
$k = 3$	0,538	**0,543**	0,523	0,449
$k = 4$	0,490	0,442	0,503	0,469
$k = 5$	0,416	0,441	0,425	0,415

Źródło: opracowanie własne.

Modele zbudowane metodą kompletnego połączenia oraz te, w których zadeklarowano identyfikowanie 5 i w dwóch przypadkach – 4 skupisk, zostały odrzucone, gdyż obliczone dla nich wartości I_S nie przekraczały 0,5. Najlepszy podział uzyskano za pomocą metody *k*-medoidów dla liczby klas równej 3. Wartość indeksu

sylwetkowego wskazuje wprawdzie słabszą strukturę klas niż w przypadku podziałów zbioru danych charakteryzowanych przez pojedyncze zmienne, jednak ze względu na to, że jest to model wielowymiarowy, uwzględniający więcej czynników demograficznych, a tym samym bardziej odpowiadający złożonej strukturze zależności w świecie rzeczywistym, zostanie on ostatecznie przyjęty do interpretacji.

Wybrany model taksonomiczny dzieli badane państwa na trzy skupiska. W klasie pierwszej znajduje się 47 państw, a w tym: kraje Afryki Środkowej, Jemen, Afganistan, Tadżykistan oraz Papua-Nowa Gwinea. Do klasy drugiej model zaklasyfikował 79 obiektów, a mianowicie kraje Afryki Północnej i Południowej, Ameryki Południowej i Środkowej oraz część państw Azji. Do klasy trzeciej trafiły 54 państwa, w tym cała Europa, Ameryka Północna oraz Australia, Nowa Zelandia, Rosja, Japonia, Południowa Korea i Urugwaj. Ze względu na relatywnie dużą liczbę obiektów zamiast wymieniać nazwy wszystkich państw w poszczególnych klasach, posłużono się mapą świata (rys. 1). Klasę pierwszą przedstawiono na mapie ciemniejszym kolorem, drugą – jaśniejszym, a trzecią jako obszar zakreskowany.



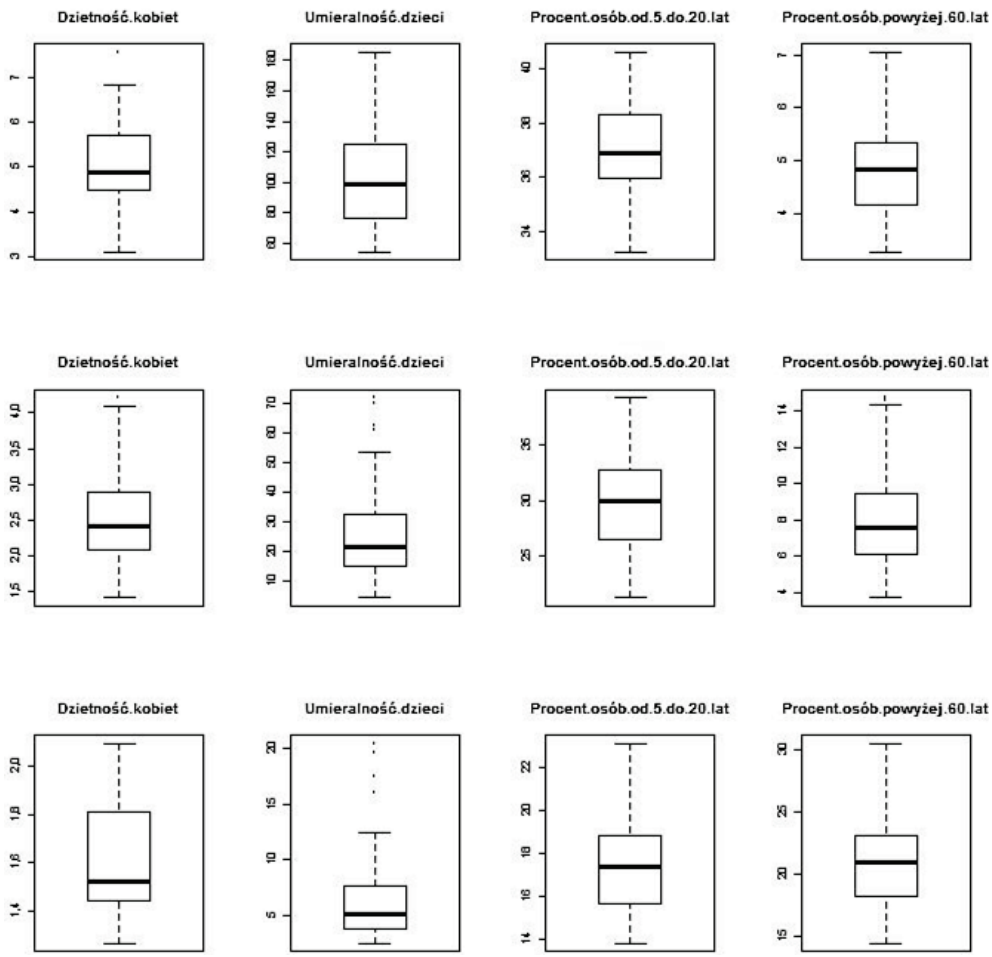
Rys. 1. Wyniki grupowania państw metodą *k*-medoidów dla $k = 3$

Źródło: opracowanie własne.

Charakterystyki tych klas przedstawiono, wykorzystując wykresy pudełkowe (rys. 2). Z rysunku 2 można odczytać, że kraje przynależące do pierwszej klasy charakteryzują się współczynnikiem dzietności równym niecałe 5. Na 1000 urodzonych dzieci umiera tam przed ukończeniem 5. roku życia około 100. W państwach z tej grupy odsetek osób w przedziale wiekowym $\langle 5,20)$ jest równy 37, natomiast procent osób w wieku powyżej 60 lat wynosi niecałe 5.

Drugie skupienie tworzą kraje, w których: dzietność kobiet jest równa około 2,5, przed ukończeniem 5 lat umiera 21 dzieci na 1000, osób w wieku od 5 do 20 lat jest 30%, natomiast w wieku powyżej 60 lat – tylko niecałe 8%.

Do trzeciej klasy należą państwa, w których liczba urodzonych dzieci, przypadających na jedną kobietę, wynosi średnio tylko 1,5. Kraje te charakteryzują się bardzo niską umieralnością dzieci przed piątym rokiem życia – 5 dzieci na 1000. Odsetek populacji w przedziale wiekowym $\langle 5,20)$ jest równy 17,5, natomiast osób powyżej 60, roku życia jest ponad 20%.



Rys. 2. Wykresy pudełkowe dla zmiennych charakteryzujących państwa:
pierwszego skupienia, oznaczonego na mapie ciemniejszym kolorem (w pierwszej linii),
drugiego skupienia, oznaczonego jaśniejszym kolorem (w drugiej linii)
oraz trzeciego skupienia, na mapie przedstawionego jako zakreskowany obszar (w trzeciej linii)

Źródło: opracowanie własne.

5. Podsumowanie

W pracy przeprowadzono analizę taksonomiczną 180 państw pod względem czynników demograficznych. Początkowo w zbiorze danych kraje były charakteryzowane przez 11 zmiennych. Jednak z uwagi na to, że wykorzystując różne warianty modeli taksonomicznych, uzyskiwano zawsze bardzo słabą strukturę klas, dokonano eliminacji cech i do ostatecznej analizy wybrano tylko te, które miały istotny wpływ na poprawę jakości grupowania. Przeprowadzone analizy pokazały, że najlepsze wyniki uzyskano, gdy państwa charakteryzowane były przez 4 zmienne:

X_1 – współczynnik dzietności kobiet,

X_3 – umieralność dzieci do 5. roku życia (na 1000 urodzeń),

X_8 – procent ludności w przedziale wiekowym $\langle 5,20)$,

X_{11} – procent ludności w wieku powyżej 60 lat.

Dla tak określonego zbioru danych najlepszą, zgodnie z przyjętymi wartościami progowymi, poważną strukturę klas odkryła metoda *k*-medoidów, dzieląc badane państwa na 3 skupiska.

Do pierwszej grupy, która charakteryzuje się wysokim współczynnikiem dzietności, lecz również wysoką umieralnością dzieci, należą kraje, które potocznie nazywamy „słabiej rozwiniętymi". W krajach tych obserwujemy strukturę ludności według wieku, w której udział ludzi młodych (od 5 do 20 lat) jest zdecydowanie wyższy, niż udział osób, które ukończyły 60 lat. Zatem do tego skupiska należą państwa, których społeczeństwa są stosunkowo „młode", a ich przyrost naturalny wysoki.

Dzietność kobiet w krajach zaliczonych do grupy drugiej jest zdecydowanie niższa niż w grupie pierwszej. Znacząco niższa jest również umieralność dzieci. Natomiast struktura ludności według wieku jest podobna jak w krajach grupy pierwszej. Tutaj również odsetek młodych osób w społeczeństwie jest kilkakrotnie wyższy niż odsetek ludzi po 60. roku życia.

Grupę trzecią tworzą kraje przede wszystkim Europy i Ameryki Północnej, często nazywane „lepiej rozwiniętymi", które w ostatnim czasie borykają się z takimi problemami demograficznymi, jak starzenie się społeczeństwa, czy spadek dzietności kobiet. W państwach tej grupy przeciętna dzietność kobiet (1,5) powoduje spadek wielkości populacji, jak również wpływa na kształtowanie się specyficznej struktury ludności według wieku, w której udział osób starszych jest równy lub większy niż udział osób młodych.

Podział na te trzy grupy można również próbować wyjaśnić, odwołując się do jednej z podstawowych koncepcji procesów demograficznych, nazywanej teorią przejścia demograficznego.

Przejście demograficzne oznacza specyficzny, historyczny proces zmian reprodukcji ludności związany z modernizacją społeczeństw [Okólski 2005]. Prowadzi

ono do zastąpienia tradycyjnego sposobu reprodukcji ludności sposobem nowoczesnym, w którym następuje radykalne obniżenie współczynnika urodzeń i współczynnika zgonów. W wyniku tego procesu zmienia się struktura ludności według wieku, a w dalszej perspektywie społeczeństwo się starzeje. Przeobrażają się również wzorce rozrodczości, zmniejsza się dzietność i najczęściej zmienia się model rodziny.

W przedstawionym badaniu kraje opisywane są tylko przez cztery charakterystyki demograficzne, jednak na podstawie otrzymanych wyników można powiedzieć, że obiekty ze skupiska trzeciego to kraje po przejściu demograficznym. Według przewidywań niektórych demografów w grupie drugiej znajdują się kraje, w których proces przejścia demograficznego jeszcze się nie skończył [zob. Okólski 2005]. Natomiast w krajach w grupie pierwszej proces ten prawdopodobnie jeszcze się nie rozpoczął.

Polska, w której w 2012 roku dzietność kobiet wynosiła 1,4, umieralność dzieci była na poziomie 5,8 (na 1000), odsetek osób w przedziale wiekowym $\langle 5,20)$ był równy 16,3%, natomiast procent osób w wieku powyżej 60 lat wynosił 19,4, zaklasyfikowana została (wraz z np. innymi krajami Europy) do skupiska trzeciego.

Literatura

Defays D. (1977), *An efficient algorithm for a complete link method*, „The Computer Journal" (British Computer Society), 20 (4), s. 364-366.

Guyon I., Gunn S., Nikravesh M., Zadeh L. (red.) (2006), *Feature Extraction, Foundations and Applications*, Springer.

Holzer J.Z. (2003), *Demografia*, Polskie Wydawnictwo Ekonomiczne PWE, Warszawa.

Jajuga K. (1993), *Statystyczna analiza wielowymiarowa*, Wydawnictwo Naukowe PWN, Warszawa.

Kauffman L., Rousseeuw P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley & Sons, New Jersey.

Kopczewska K., Kopczewski T., Wójcik P. (2009), *Metody ilościowe w R. Aplikacje ekonomiczne i finansowe*, CeDeWu, Warszawa.

MacQueen J.B. (1967), *Some Methods for Classification and Analysis of Multivariate Observations*, „Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability", 1, University of California Press, s. 281-297.

Okólski M. (2005), *Demografia. Podstawowe pojęcia, procesy i teorie w encyklopedycznym zarysie*, Wydawnictwo Naukowe Scholar, Warszawa.

Pociecha J., Podolec B., Sokołowski A., Zając K. (1986), *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa.

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

Ward J.H. Jr. (1963), *Hierarchical Grouping to Optimize an Objective Function*, „Journal of the American Statistical Association", 58, s. 236-244.

CLUSTER ANALYSIS OF COUNTRIES WITH RESPECT TO FERTILITY RATE AND OTHER DEMOGRAPHIC FACTORS

Summary: The paper presents cluster analysis of 180 countries. In the first stage, countries were described by 11 demographic variables. Using all the variables led to a poor class structure, thus the procedure for variables selection was performed in the next stage. Only fertility rate, children mortality rate, population aged 5–20 (% of total) and population aged 60 and older (% of total) had a significant impact on the clustering quality. The best model was built using *k*–medoids. As a result the countries were grouped into three clusters.

Keywords: cluster analysis, demography, variable selection.