PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Artur Mikulec
Uniwersytet Łódzki

WYBRANE METODY KLASYFIKACJI DLA DUŻYCH BAZ DANYCH W ANALIZIE DEMOGRAFICZNEGO STARZENIA SIĘ LUDNOŚCI W KRAJACH UE I EFTA*

1. Wstęp

Do podstawowych kryteriów decydujących o możliwości zastosowania danej metody analizy skupień należą m.in. skala pomiaru wykorzystywanych zmiennych i konkretny cel analizy [Kaufman, Rousseeuw 1990, 2005]. Uniwersalność metod sprawia, że analiza ich własności oraz wskazówki dotyczące możliwości ich aplikacji z reguły nie pozwalają na wybór jednej „ścieżki poszukiwania rozwiązania", a duża liczba dopuszczalnych oraz możliwych do zaakceptowania rozwiązań najczęściej nie upraszcza analizy.

Spośród dużej liczby metod analizy skupień w badaniach społeczno-ekonomicznych o charakterze regionalnym często i chętnie wykorzystywane są metody aglomeracyjne i podziałowe (optymalizacyjne), a w szczególności metody Warda (*Ward*) i *k*-średnich (*k-means*) oparte na idei środka ciężkości obiektów w skupieniach.

Celem referatu jest prezentacja algorytmów wybranych metod podziałowych analizy skupień, tj. PAM (*Partitioning Around Medoids*), CLARA (*Clustering LARge Applications*) i CLARANS (*Clustering Large Applications based on RANndomised Search*), wykorzystujących do opisu poszczególnych skupień obiekt położony najbliżej środka ciężkości (*k-medoids*), a nie sam środek ciężkości. Wymienione algorytmy przeznaczone są do analizy „dużych baz danych", jednak od momentu ich powstania zmianie uległa definicja tego typu bazy danych. Rozwój techniki spowodował zwiększenie możliwości obliczeniowych komputerów, powodując podniesienie granicy skutecznego i efektywnego wykorzystania tych algorytmów jako narzędzia analizy z jednoczesnym zwiększeniem wielkości gromadzonych i profesjonalnie przetwarzanych baz danych.

* Praca naukowa finansowana ze środków na naukę w latach 2008-2009 jako projekt badawczy nr N N111 436734.

Algorytmy PAM i CLARA zastosowano do analizy demograficznego starzenia się ludności w krajach UE i EFTA – objętych wspólną otwartą metodą koordynacji (OMK) systemów emerytalnych. Do obliczeń wykorzystano dane statystyczne o liczbie ludności według wieku na poziomie NTS2 krajów UE i EFTA, a uzyskane wyniki za pomocą indeksu sylwetkowego (*Silhouette Index,* SI) porównano pod względem jakości grupowania z klasycznym algorytmem podziałowym *k*-średnich (Hartigana-Wonga).

2. Opis działania algorytmów PAM, CLARA i CLARANS

Algorytm PAM opracowany przez Kaufmana i Rousseeuwa [Kaufman, Rousseeuw 1990, 2005] polega na wyszukiwaniu w analizowanym zbiorze *n* obiektów *k* reprezentantów (obiektów) leżących najbliżej centrów skupień, którzy razem z pozostałymi przypisanymi do nich według kryterium najbliższej odległości obiektami utworzą poszukiwane skupienia. Tak rozumiane środki ciężkości skupień (obiekty) zostały przez autorów nazwane medoidami (*medoids*). Szczegółowy schemat działania algorytmu PAM można zapisać następująco [Ng, Han 1994; Han, Kamber 2000; Kolenda 2006]:

Krok 1. Wybierz arbitralnie lub wylosuj *k* obiektów, czyli potencjalnych centrów skupień.

Krok 2. Przypisz wszystkie pozostałe obiekty do najbliższych im centrów skupień.

Krok 3. Wyznacz funkcję kosztu TC_{ih} dla każdej pary obiektów (i,h)[1], w której jeden jest reprezentantem, a drugi nie (w wariancie PAM II algorytmu funkcją celu może być suma odległości wszystkich obiektów od centrów skupień).

Krok 4. Zmień pojedynczy obiekt będący reprezentantem skupienia (*medoids*) na inny niebędący reprezentantem dla tej pary (i,h), dla której wartość TC_{ih} jest najmniejsza, tzn. ujemna (w wariancie PAM II, jeśli następuje zmniejszenie wartości funkcji celu).

Krok 5. Powtarzaj kroki 2-4 do momentu aż nie nastąpi dalsza zmiana reprezentanta, tzn. gdy wartość TC_{ih} nadal jest ujemna (w wariancie II, gdy następuje zmniejszenie funkcji celu, tj. sumy odległości pomiędzy wszystkimi obiektami i ich centrami skupień), w przeciwnym razie -> STOP.

Algorytm CLARA autorstwa Kaufmana i Rousseeuwa jest dwuetapowy [Kaufman, Rousseeuw, 1990, 2005; Han, Kamber 2000]. Najpierw z analizowanego zbioru danych losowana jest próba $40 + 2k$ obiektów, która jest dzielona na *k* skupień z wykorzystaniem metody PAM. W wyniku obliczeń wyznaczonych zostaje *k* reprezentantów skupień, do których w kroku drugim przypisywane są wszystkie pozostałe obiekty niebędące w próbie. Miarą jakości grupowania jest średnia odległość wszystkich analizowanych obiektów w stosunku do wyodrębnionych reprezentantów grup. Schematycznie działanie algorytmu CLARA można zapisać w postaci poniższych kroków [Ng, Han 1994; Han, Kamber 2000]:

[1] Funkcja kosztu TC_{ih} jest sumą odległości wszystkich obiektów niebędących reprezentantami skupień od najbliższych im obiektów-reprezentantów obliczana po każdej pojedynczej zamianie dwóch dowolnych obiektów O_i, tj. reprezentanta z obiektem O_h, który nie należy do zbioru reprezentantów skupień.

Krok 1. Użytkownik definiuje liczbę k poszukiwanych skupień.

Krok 2. Dla $i = 1$ do 5 (liczba powtórzeń całego algorytmu).

Krok 3. Pobierz losowo próbę $40+2k$ obiektów ze zbioru danych (dla $i = 2,\ldots,5$ algorytm losuje $40+k$ obiektów, gdyż k jest zapamiętywane z poprzedniego rozwiązania) i dla tak wylosowanej próby zastosuj algorytm PAM, aby wyznaczyć k medoidów.

Krok 4. Wszystkie inne analizowane obiekty przypisz do najbliższych im k wyznaczonych reprezentantów (centrów) skupień.

Krok 5. Dla pozostałych $40+k$ obiektów z próby (dla $i = 1,\ldots,5$) określ funkcję kosztu TC_{ih}. Jeśli obecna wartość funkcji kosztu (TC_{ih}) jest mniejsza niż ta z poprzedniego losowania (dla $i = 1$, $TC_{ih} = 0$), to ustal ją jako bieżące minimum, zapamiętaj wyznaczonych k reprezentantów i wróć do kroku 3.

Krok 6. Przejdź do kroku 2 i rozpocznij kolejną iterację lub -> STOP[2].

Ostatni ze wspomnianych algorytmów CLARANS jest kombinacją algorytmu PAM i metody próbkowania zbioru danych. W celu zrozumienia istoty działania algorytmu należy zdefiniować pojęcie węzła oraz sąsiada. Węzłem jest dowolny zbiór poszukiwanych k medoidów, natomiast sąsiadami są dwa węzły różniące się tylko jednym obiektem, inaczej mówiąc, zawierające k-1 tych samych obiektów. Z definicji tych wynika, że każdy węzeł ($G_{n,k}$) ma maksymalnie $k(n-k)$ sąsiadów (S). „Z przejściem" pomiędzy węzłami, tj. potencjalnymi rozwiązaniami k medoidów, związany jest omawiany wcześniej koszt zmiany jednego reprezentanta TC_{ih}. Idea algorytmu, po wstępnym wyborze k reprezentantów – węzła, polega na znalezieniu rozwiązania (węzła, zbioru k medoidów) wśród sąsiadów, które zminimalizuje wartość funkcji kosztu TC_{ih}.

Pierwotnie algorytm PAM był dedykowany analizie danych co najwyżej $k = 5$ skupień w zbiorze $n = 100$ obiektów, CLARA była rekomendowana do poszukiwania $k = 10$ skupień w zbiorze $n = 1000$ obiektów, natomiast CLARANS − do analizy jeszcze większych baz danych [Ng, Han 1994]. Przy obecnym postępie techniki obliczeniowej podane granice „efektywności" tych algorytmów należy traktować umownie.

3. Analiza demograficznego starzenia się ludności

Obszar UE 27 w celach statystycznych jest podzielony na poziomie NTS2 na 271 „regionów", a w obszar krajów EFTA składa się z 17 tego typu „regionów". W referacie rozpatrywano zagadnienie demograficznego starzenia się ludności w krajach UE i EFTA, tj. zmianę stopnia starości społeczeństwa w czasie na poziomie NTS2[3]. Ze względu na braki danych statystycznych dla niektórych krajów UE na

[2] Efektywność CLARA zależy od rozmiaru próby, która powinna być wartością $min\{n; 40+2k\}$, gdzie n to liczba analizowanych obiektów. Satysfakcjonujące rozwiązanie uzyskuje się już przy warunku pięciokrotnego powtórzenia algorytmu ($i = 5$). Warto zauważyć, że w krokach 3-6 CLARA wielokrotnie próbkuje cały analizowany zbiór, poszukując „najlepszego rozwiązania", dając w wyniku k reprezentantów skupień oraz informacji o przynależności wszystkich analizowanych obiektów do poszczególnych reprezentantów (krok 3).

[3] Kraje EFTA to Islandia, Liechtenstein, Norwegia i Szwajcaria. Informacje na temat nomenklatury NTS dostępne są pod adresem: http://ec.europa.eu/eurostat/ramon/nuts/introduction_regions_en.html.

poziomie NTS2 ostatecznie do obliczeń wykorzystano dane statystyczne o liczbie ludności według wieku za lata 2000 i 2004 dla 273 „regionów" krajów UE i EFTA.

Wyznaczono wskaźniki demograficznego starzenia się ludności [Rosset 1959; Frątczak 1984; Długosz 1998; Mikulec 2007], z których po analizie zmienności i korelacji pozostawiono trzy. Względny wskaźnik demograficznego starzenia się ludności w przeliczeniu na 1000 mieszkańców – modyfikacja wskaźnika zaproponowanego przez Długosza – jest następujący:

$$W_{sd} = \frac{[U_{(0-14)\,t} - U_{(0-14)\,t+n}] + [U_{(>65)\,t+n} - U_{(>65)\,t}]}{[U_{(og)\,t} + U_{(og)\,t+n}]/2} * 1000, \qquad (1)$$

gdzie: $U_{(0-14)\,t}$, $U_{(0-14)\,t+n}$ oznacza udział ludności w wieku 0-14 lat odpowiednio na początku i końcu badanego okresu, $U_{(>65)\,t+n}$, $U_{(>65)\,t}$ to udział ludności w wieku powyżej 65 lat odpowiednio na końcu i na początku badanego okresu, a $[U_{(og)\,t} + U_{(og)\,t+n}]/2$ to średnia liczba ludności ogółem na początku i na końcu badanego okresu. W_{sd} ukazuje stopień starzenia się ludności, a im wyższa od zera jest jego wartość, tym proces starzenia się jest silniejszy. Wartości W_{sd} niższe od zera oznaczają odmładzanie się społeczeństwa. Wskaźnik postępu starzenia się ludności opiera się na przyroście ludności w wieku 60/65 lat i więcej w stosunku do przyrostu ludności ogółem i wyrażony jest w procentach oraz opiera się na ogólnym współczynniku obciążenia demograficznego, tj. średnim rocznym przyroście współczynnika obciążenia demograficznego w analizowanym okresie [Frątczak 1984].

Z bazy danych dla 273 „regionów" na podstawie analizy wykresów rozrzutu wartości wskaźników wyłączono 8 obiektów, które uniemożliwiały dokonanie nietrywialnego podziału zbioru obiektów na jednorodne grupy według kolejności: DE50 (Bremen), DE30 (Berlin), UKL1 (West Wales and The Valleys), UKG2 (Shropshire and Staffordshire), SE21 (Småland med öarna), GR11 (Anatoliki Makedonia, Thraki), GR13 (Dytiki Makedonia) oraz RO21 (Nord-Est) w Rumunii. Można przyjąć, że są to „regiony" o szczególnie wysokim tempie demograficznego starzenia się ludności. Dalszą analizę przeprowadzono na podstawie 265 „regionów" krajów UE i EFTA.

4. Porównanie wyników algorytmów *k*-średnich, PAM oraz CLARA

Uwzględniając liczbę grupowanych obiektów (n = 265), do analizy zastosowano algorytmy PAM i CLARA, a ich wyniki porównano z klasycznym algorytmem *k*-średnich (Hartigana-Wonga). W obliczeniach wykorzystano pakiet `cluster` (autorstwa Rousseeuwa i in.), `clusterSim` (którego autorami są Walesiak i Dudek) oraz pakiet `stats` środowiska R. Nie było przesłanek odnośnie do poszukiwanej liczby *k* preferowanych skupień.

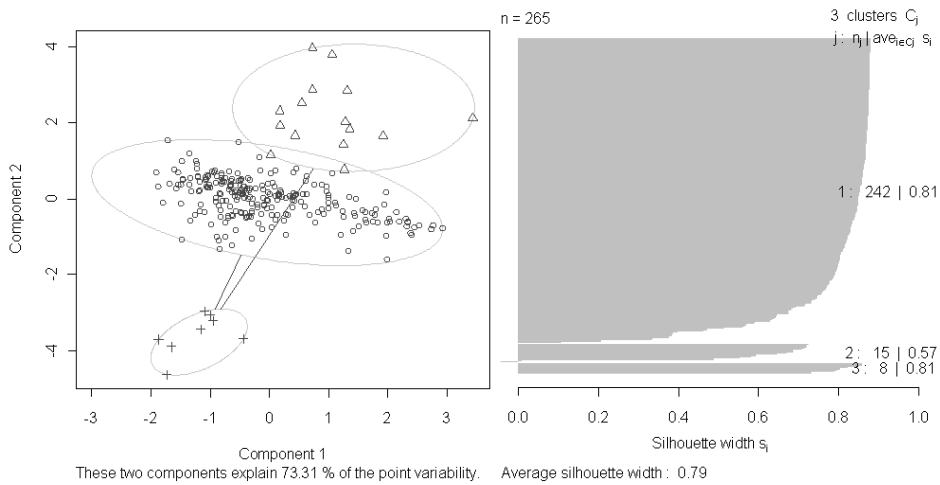
Wykonano w pętli po 1000 powtórzeń algorytmów metody *k*-średnich i PAM z losowym doborem punktów startowych dla liczby skupień $s = 2,\ldots,5$, za każdym razem oceniając za pomocą indeksu sylwetkowego *SI* wynik uzyskanego grupowania. Zastosowanie właśnie takiej kolejności obliczeń, tzn. grupowanie dla różnej liczby skupień od 2 do 5, w ramach (wewnątrz) 1000 powtórzeń każdego z algorytmów pozwoliło (dzięki wykorzystaniu indeksu *SI*) odpowiedzieć na pytanie: jak często w pojedynczym przebiegu algorytmu „wygrywają rozwiązania podziału obiektów" dla poszczególnych *k* w zakresie od 2 do 5 skupień. Możliwe było także określenie parametru *k* „najlepszego" uzyskanego podziału zbioru obiektów. Najlepsze pojedyncze rozwiązanie uzyskane za pomocą algorytmu *k*-średnich ($SI = 0{,}823$) zawierało 2 skupienia obiektów. Warto jednak zwrócić uwagę na częstość występowania poszczególnych rozwiązań. Wśród 1000 końcowych podziałów metody *k*-średnich 2 skupienia wystąpiły 740 razy, a wynik grupowania dla 3 skupień „wygrał" 260 razy. Algorytm PAM dał w wyniku „najlepszego rozwiązania" 3 skupienia „regionów" ($SI = 0{,}793$), a na 1000 powtórzeń: 2 skupienia wystąpiły 147 razy, a 3 skupienia − 732 razy. W przypadku PAM uzyskano także wyniki dla 5 skupień (121 razy). Czas obliczeń algorytmów wyniósł odpowiednio 7:06 minut (*k*-średnich) i 5:23 minut (PAM).

Algorytm CLARA, bazujący na wielokrotnym losowaniu próby ze zbioru analizowanych obiektów, także został powtórzony po 1000 razy w zakresie od 2 do 5 skupień. Czas wszystkich obliczeń wyniósł 4:33 minut i pozwolił na uzyskanie wysokiej jakości grupowania. Najlepszym pojedynczym rozwiązaniem był podział na 2 skupienia uzyskany na podstawie próby 44 obiektów, dla której indeks *SI* wyniósł 0,908. Przyporządkowanie wszystkich obiektów do uzyskanych reprezentantów skupień spowodowało spadek wartość indeksu sylwetkowego do 0,824. Analizując częstość uzyskiwania rozwiązań o liczbie od 2 do 5 skupień, należy stwierdzić, że rozwiązanie z 3 skupieniami (511 razy) częściej „wygrywało" z rozwiązaniami z 2 skupieniami (479 razy), 4 skupieniami (6 razy) i 5 skupieniami (4 razy). Wartość indeksu sylwetkowego dla najlepszego podziału obiektów na 3 skupienia algorytmem CLARA dla wylosowanej próby wyniosła $SI = 0{,}883$, a dla całego zbioru obiektów $SI = 0{,}804$.

5. Wyniki analizy demograficznego starzenia się ludności

Jako wynik analizy skupień „regionów" NTS2 z punktu widzenia demograficznego starzenia się ludności uznano ostatecznie 3 skupienia uzyskane dla większej liczby powtórzeń algorytmów PAM i CLARA. Szczegółowa analizy „najlepszego" grupowania obiektów dla 3 skupień algorytmem PAM i CLARA wskazuje, że uzyskane wyniki są bardzo podobne. Jest to dość oczywiste, gdyż CLARA, analizując próbę, wykorzystuje algorytm PAM. Potwierdza się jednak duża efektywność algorytmu CLARA, który na podstawie próby 46 obiektów, stanowiącej 17,4% obiektów, zdołał „znaleźć" w krótszym czasie prawie identyczne rozwiązanie. Uzyskane skupienia w przypadku PAM składały się z {242}, {15} i {8} „regionów", a skupienia CLARA − z {244}, {13} i {8} „regionów". Szczegółowe porównanie wykazało,

że różniły się one tylko 2 obiektami (zob. rys. 1). W przypadku metod CLARA i PAM medoidy same w sobie mają użyteczną interpretację – obiektów reprezentujących skupienia – nie można ich jednak porównywać (czy są takie same), gdyż są losowane odpowiednio z próby i ze zbioru wszystkich obiektów.

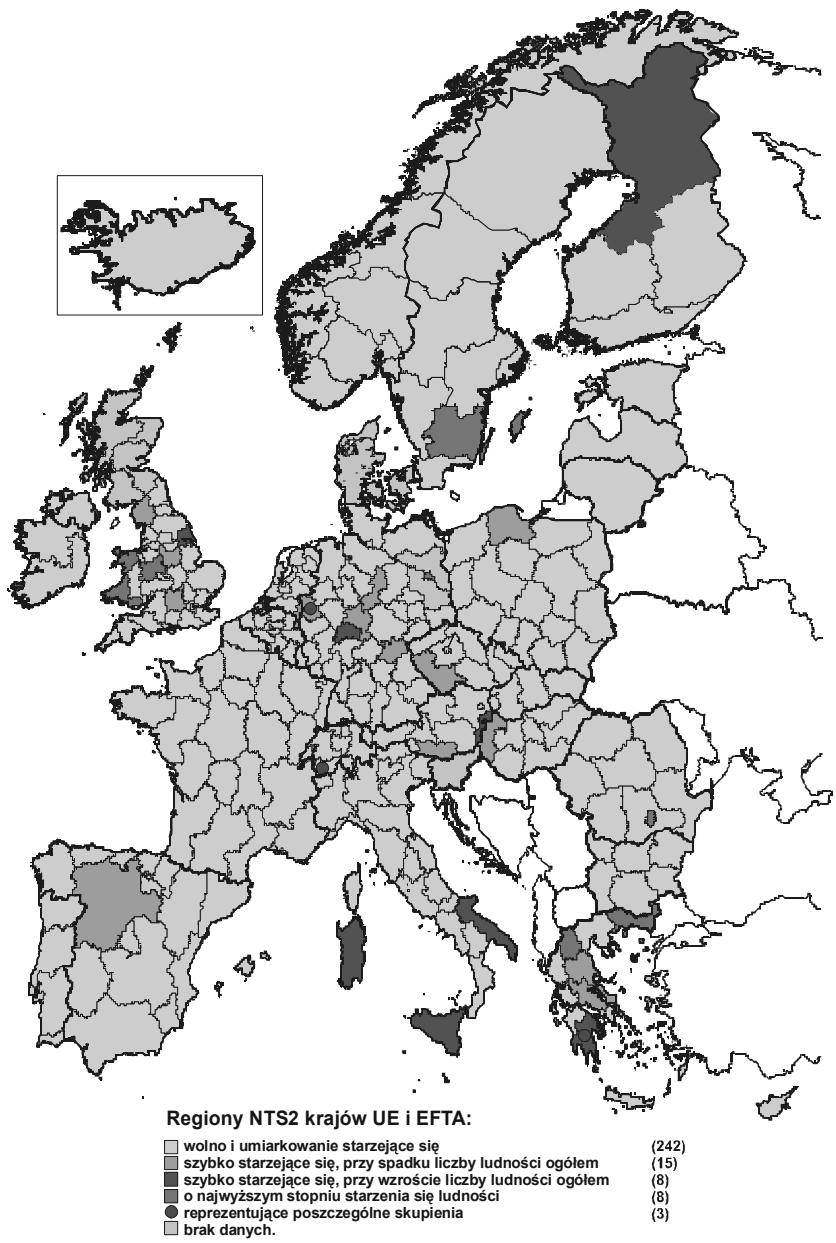


Rys. 1. Wynik analizy skupień metodą PAM (3 skupienia)
Źródło: pakiet cluster programu R na podstawie obliczeń własnych.

Przechodząc do analizy uzyskanych wyników i profilowania skupień, które przedstawione zostały na rys. 2, należy zauważyć, że reprezentantem najliczniejszego 1 skupienia o 242 „regionach" był CH01 (Région Lémanique) w Szwajcarii; najbliżej „centrum" 2 skupienia 15 „regionów" znajdował się DEA1 (Düsseldorf) w Niemczech; 3 skupienie 8 „regionów" reprezentowane było przez GR25 (Peloponnisos) w Grecji. Jako 4 skupienie można chyba uznać „regiony" wyłączone z analizy, które charakteryzowały się bardzo wysokim tempem demograficznego starzenia się ludności.

Région Lémanique w Szwajcarii reprezentujący skupienie 1 charakteryzował się najniższym wskaźnikiem demograficznego starzenia się ludności na 1000 mieszkańców − 9,6 osób (wartość średnia to 15,4 osób) − i zbliżonym do wartości średniej równej 23,2% wskaźnikiem postępu demograficznego starzenia się ludności opartym na przyroście ludności w wieku 60/65 lat i więcej w % ludności ogółem wynoszącym 27,6%. Wartość wskaźnika na bazie przyrostu ogólnego współczynnika obciążenia demograficznego wyniosła 0,014 (wartość średnia to –0,173). Regiony skupienia 1 można określić, jako „wolno i umiarkowanie starzejące się".

Düsseldorf w Niemczech skupiał 15 „regionów" o wysokim tempie demograficznego starzenia się ludności. Średnia wartość wskaźnika demograficznego starzenia się ludności na 1000 mieszkańców dla tego regionu wyniosła 22,1 osób. W Düsseldorfie występował ujemny wskaźnik postępu starzenia się ludności oparty



Rys. 2. Przestrzenne zróżnicowanie procesu demograficznego starzenia się ludności w krajach UE i EFTA
Źródło: opracowanie własne na podstawie obliczeń pakietu `cluster` środowiska R.

na przyroście ludności w wieku 60/65 lat i więcej w % ludności ogółem rzędu – 415,4%. To oznacza, że przyrostowi liczby ludności w wieku starszym towarzyszył spadek liczy ludności ogółem oraz że przyrost liczby osób starszych był 4-krotnie większy od przyrostu liczby ludności ogółem – tego typu zmiany były typowe dla wszystkich regionów w tym skupieniu. Wartość trzeciego wskaźnika dotyczącego średniej zmiany współczynnika obciążenia demograficznego wyniosła 0,722 i również wskazywała na szybkie tempo procesów starzenia się ludności.

Peloponnisos w Grecji również reprezentuje „regiony" o szybkim tempie demograficznego starzenia się ludności (skupienie 8 regionów), lecz objawiającym się: wysokim wskaźnikiem starzenia się demograficznego ludności – 21,3 osób na 1000 mieszkańców, wysokim przyrostem liczby osób starszych połączonym z przyrostem liczby ludności ogółem wynoszącym 667% (przyrost ludności w wieku 60/65+ jest 6-krotnie większy niż ludności ogółem), a także ujemną wartością (–0,656) odnoszącego się do zmian ogólnego współczynnika obciążenia demograficznego wskaźnika postępu starzenia się ludności. Jego ujemna wartość sugeruje częściowe „odmładzanie się społeczeństwa", co jest wynikiem przyrostu ludności w wieku 0-14 lat, tj. w wieku przedprodukcyjnym.

6. Podsumowanie i wnioski

W referacie zaprezentowano algorytmy PAM, CLARA i CALARANS, a dwa pierwsze zastosowano do analizy demograficznego starzenia się ludności w krajach UE i EFTA na poziomie NTS2. Na podstawie przeprowadzonych obliczeń można stwierdzić, że algorytmy PAM i CLARA (*k*-medoidów) są bardziej „konsekwentne" i tworzą bardziej stabilne skupienia w porównaniu z algorytmem *k*-średnich (Hartigana-Wonga). W przypadku analiz „regionalnych" mają też tę dodatkową zaletę, że „centra" ich skupień są analizowanymi obiektami, którym można nadać „rzeczywistą" interpretację.

W wyniku przeprowadzonych analiz przyjęto rozwiązanie podziału 265 „regionów" na 3 skupienia odpowiednio: 242 „wolno i umiarkowanie starzejących się regionów" oraz 15 i 8 „regionów o wysokim tempie starzenia się ludności", lecz nieco innym przebiegu omawianego procesu – gorsza sytuacja występuje w skupieniu 15 regionów. Obszary wyłączone z analizy o najwyższym stopniu starzenia się ludności występujące w Grecji, Niemczech, Rumunii, Szwecji i Wielkiej Brytanii można potraktować jako 4 skupienie.

Literatura

Długosz Z. (1998), *Próba określenia zmian starości demograficznej Polski w ujęciu przestrzennym*, „Wiadomości Statystyczne", 3, s.15-27.

Frątczak E. (1984), *Proces starzenia się ludności Polski a proces urbanizacji*, Monografie i Opracowania nr 157, SGH ISID, Warszawa.

Han J., Kamber M. (2000), *Data mining: concept and techniques*, Morgan Kaufmann Publishers.

Kaufmann L., Rousseeuw P.J. (1990, 2005), *Finding groups in data. An introduction to cluster analysis*, Wiley-Interscience, New York.
Kolenda M. (2006), *Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych*, AE, Wrocław.
Mikulec A. (2007), *Analiza starzenia się ludności w polskich podregionach*, „Wiadomości Statystyczne", 1, s. 62-75.
Ng R., Han J. (1994), *Efficient and effective clustering methods for spatial data mining*, Proceedings of the 20th VLDB Conference, Chile.
Rosset E. (1959), *Proces starzenia się ludności*, PWG, Warszawa.

CHOSEN CLUSTERING METHODS FOR LARGE DATA SETS IN ANALYSIS OF POPULATION'S AGEING IN THE EU AND EFTA COUNTRIES

Summary

In the study, chosen partitioning clustering methods have been discussed – the PAM algorithm (*Partitioning Around Medoids*) – Kaufman, Rousseeuw and also clustering methods for large data sets, the CLARA algorithm (*Clustering LARge Applications*) Kaufman, Rousseeuw and the CLARANS algorithm (*Clustering Large Applications based on RANndomised Search*) – Ng, Han.

The PAM and CLARA algorithms were applied to the analysis of population's ageing of society in the EU and EFTA countries. In the calculations, statistics data of population by age at NUTS-2 level for the EU and EFTA countries were applied. Obtained results were compared with results of clustering based on the k-means algorithm (*Hartigan-Wong*).