PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Piotr Łukasiewicz, Grzegorz Koszela
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Arkadiusz Orłowski
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie, Instytut Fizyki PAN w Warszawie

KLASYFIKACJA ROZKŁADÓW DOCHODÓW GOSPODARSTW DOMOWYCH

1. Wstęp

W badaniach społecznych dochody ludności są najważniejszym czynnikiem odzwierciedlającym poziom życia ludności oraz jego zróżnicowanie. Zróżnicowanie to wyraźnie ujawnia się w różnych podziałach – zarówno terytorialnych, jak i społeczno-ekonomicznych, np. według głównego źródła dochodu, wykształcenia, typu biologicznego rodziny i innych. Najczęściej porównania poziomu dochodów dokonuje się, budując rankingi według dochodu średniego. Oczywiście nie uwzględnia się wówczas innych ważnych informacji dotyczących rozkładu cechy, takich jak np. stopień zróżnicowania czy koncentracja dochodu. Biorąc pod uwagę kilka charakterystyk rozkładu, jak wartość średnia, mediana, odchylenie standardowe, współczynnik asymetrii, współczynnik koncentracji i inne, do porównania rozkładów można wykorzystać znane metody klasyfikacyjne. Ten sposób postępowania przedstawiony został w pracy [Łukasiewicz, Orłowski 2003a] i wykorzystany następnie w pracy [Łukasiewicz, Orłowski 2003b]. Problemem okazał się wybór odpowiednich charakterystyk rozkładu, a także pytanie, w jakim stopniu charakterystyki te odzwierciedlają cały rozkład dochodu. W pracy [Łukasiewicz i in. 2007] przedstawiono metodę klasyfikacji rozkładów opartą na konstrukcji teoretycznego rozkładu dochodu. Prezentowane wyniki są kontynuacją tych badań. W pracy autorzy koncentrują się na aspekcie metodycznym, nie dokonują ocen i porównań sytuacji ekonomicznej ludności. Wykorzystywane są dane z 2000 r. oraz modele dochodów skonstruowane we wcześniejszych badaniach.

Pierwszym celem pracy jest przedstawienie metody klasyfikacji rozkładów dochodów z wykorzystaniem rozkładu teoretycznego (modelu dochodu) oraz wybra-

nych metod taksonomicznych. Drugi cel to wyznaczenie najlepszego podziału rozkładów dochodów za pomocą miar zwartości i odrębności skupień.

2. Materiał empiryczny

Dane wykorzystane w pracy to indywidualne informacje o dochodzie rozporządzalnym gospodarstw domowych, pochodzące z badań budżetów gospodarstw domowych za rok 2000. W badaniach tych, prowadzonych corocznie przez GUS, stosowany jest terytorialny, warstwowy i dwustopniowy schemat doboru próby, odpowiedni do prowadzenia analiz zarówno w przekroju społeczno-ekonomicznym ludności kraju, jak i w przekroju terytorialnym [*Budżety gospodarstw...* 2001]. Ze zbioru 36 163 gospodarstw usunięto te jednostki, które wykazały w swoich sprawozdaniach ujemny lub zerowy dochód. Otrzymano w ten sposób zbiór danych o liczebności 35 951 gospodarstw, który podzielono według województw na 16 prób. Rozpatrywaną w pracy kategorią dochodu jest roczny dochód gospodarstwa w przeliczeniu na osobę. Każda z wyodrębnionych prób stanowi empiryczny rozkład dochodów gospodarstw domowych dla danego województwa. Przeprowadzono test zgodności Kołmogorowa-Smirnowa [Plucińska, Pluciński 2000] dla wszystkich par rozkładów empirycznych. Test rozstrzyga, czy badane rozkłady empiryczne pochodzą z tego samego rozkładu w populacji (hipoteza zerowa). Spośród 120 par rozkładów w 105 przypadkach test wskazuje na odrzucenie hipotezy zerowej. Można więc stwierdzić, że badana populacja gospodarstw domowych nie jest jednorodna, a poszukiwanie podobnych rozkładów jest celowe.

3. Model rozkładu dochodów

W pracy [Łukasiewicz, Orłowski 2003c] wykazano, że bardzo dobry opis rozkładu dochodów gospodarstw domowych można uzyskać za pomocą dwóch trzyparametrycznych modeli: rozkładu Daguma i rozkładu Singha-Maddali. Oba modele charakteryzują się równie wysokim poziomem dopasowania do danych empirycznych. Do niniejszych badań wybrano rozkład Daguma (rozkład Burra III). Model ten jest często wykorzystywany do matematycznego opisu rozkładu dochodów. Ma tę przewagę nad modelem Singha-Maddali, że skonstruowany został na podstawie pewnych przesłanek o charakterze ekonomicznym, dotyczących ogólnych własności rozkładu dochodów (por. [Dagum 1977]). Z polskich autorów model Daguma badali i stosowali m.in.: Jędrzejczak [1990], Kośny [2001], Łukasiewicz, Orłowski [2003 a, b, c]. Natomiast model Singha-Maddali wykorzystywał S.M. Kot do opisu rozkładu płac [*Analiza ekonometryczna...* 1999]. Funkcja gęstości rozkładu Daguma określona jest wzorem:

$$f(x) = \frac{abc}{x^{b+1}(1+ax^{-b})^{c+1}}, \tag{1}$$

gdzie $a > 1$, $b > 1$, $c > 0$ są parametrami modelu.

Do estymacji parametrów modelu (1) zastosowano metodę największej wiarygodności opartą na danych indywidualnych $x_1, x_2, ..., x_n$. Metoda ta polega na wyznaczeniu takich ocen $\hat{a}$, $\hat{b}$, $\hat{c}$ parametrów funkcji gęstości f, dla których funkcja wiarygodności L, określona wzorem:

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \tag{2}$$

osiąga maksimum. W powyższym wzorze $f(x_i \mid \theta)$ oznacza wartość funkcji gęstości w punkcie x_i $(i = 1, 2, ..., n)$, a $\theta = [a, b, c]$ jest wektorem szacowanych parametrów funkcji f.

W celu oceny jakości dopasowania oszacowanego rozkładu teoretycznego do rozkładu empirycznego wykorzystano miarę ρ^2, która jest równa kwadratowi współczynnika korelacji ρ między kwantylami empirycznymi i teoretycznymi [*Analiza ekonometryczna...* 1999]. W każdym z rozpatrywanych przypadków liczbę ρ wyznaczono, wykorzystując dane indywidualne i biorąc pod uwagę kwantyle rzędu $\frac{r}{200}$, gdzie $r = 1, 2, ..., 199$. Dla każdego z 16 województw oszacowano – opisaną wyżej metodą – model dochodów (1). Stopień zgodności poszczególnych modeli z rozkładami empirycznymi jest bardzo wysoki, jedynie w przypadku województw lubuskiego i podkarpackiego wartość ρ^2 jest nieco niższa od 0,99. Wyniki estymacji przedstawiono w tab. 1.

Tabela 1. Oceny parametrów i dopasowanie oszacowanych modeli dochodów

Nr	Województwo	$\hat{a}$	$\hat{b}$	$\hat{c}$	ρ^2
1	Dolnośląskie	941,52	3,3175	0,8061	0,9969
2	Kujawsko-pomorskie	1305,05	3,5017	0,7256	0,9975
3	Lubelskie	2840,68	3,8364	0,5397	0,9958
4	Lubuskie	1440,72	3,5718	0,8407	0,9857
5	Łódzkie	2624,12	3,6620	0,7174	0,9971
6	Małopolskie	1019,02	3,3950	0,8333	0,9991
7	Mazowieckie	506,74	2,8542	0,9249	0,9997
8	Opolskie	1655,24	3,4977	0,6981	0,9962
9	Podkarpackie	1065,29	3,5937	0,7203	0,9824
10	Podlaskie	2855,49	3,7911	0,5749	0,9984
11	Pomorskie	862,12	3,1757	0,7688	0,9983
12	Śląskie	2174,38	3,6157	0,8365	0,9971
13	Świętokrzyskie	3197,74	3,9289	0,5604	0,9970
14	Warmińsko-mazurskie	971,85	3,4817	0,7390	0,9992
15	Wielkopolskie	777,05	3,2938	0,8044	0,9995
16	Zachodniopomorskie	483,62	3,0758	0,8796	0,9929

Źródło: obliczenia własne, [Łukasiewicz i in. 2007].

4. Klasyfikacja rozkładów dochodów

Oszacowane modele dochodów potraktowane zostały jako obiekty klasyfikacji. Każdy obiekt scharakteryzowany jest przez trzy cechy reprezentowane przez zmienne $\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$

($i = 1, 2, ..., 16$) i może być uważany za punkt w 3-wymiarowej przestrzeni. Zmienne poddano standaryzacji, a macierz odległości między obiektami zbudowano, wykorzystując metrykę euklidesową. Do grupowania obiektów zastosowano pięć metod taksonomicznych: 1) metodę *k*-średnich, 2) metodę najdalszego sąsiedztwa, 3) metodę mediany, 4) metodę średnich połączeń, 5) metodę Warda.

Metoda *k*-średnich należy do grupy metod podziałowych, optymalizacyjno-iteracyjnych. Pozostałe metody zaliczane są do grupy aglomeracyjnych metod hierarchicznych. Teoretycznie, przy grupowaniu n obiektów, liczba klas może być liczbą całkowitą z przedziału $[1; n]$. Jednak rozpatrywanie zbyt dużej liczby klas prowadzi do powstawania skupień jednoelementowych, co z praktycznego punktu widzenia nie jest celowe. Z tego względu w badaniu przyjęto maksymalną liczbę klas równą $n/2$ (średnio dwa obiekty na jedno skupienie). Rozpatrywano więc wszystkie podziały 16 obiektów (rozkładów dochodów), jakie uzyskano za pomocą wymienionych metod taksonomicznych na 2, 3, 4, 5, 6, 7 i 8 klas. Wyniki klasyfikacji przedstawiono w dalszej części opracowania.

5. Miary jakości klasyfikacji

Do oceny jakości otrzymanych podziałów wykorzystano mierniki zwartości (homogeniczności) i odrębności (heterogeniczności) skupień. Niech dany będzie podział obiektów $\gamma_1, \gamma_2, ..., \gamma_{16}$ na k skupień $\Gamma_1, \Gamma_1, ..., \Gamma_k$ o liczebnościach $n_1, n_2, ..., n_k$ ($n_1 + n_2 + ... + n_k = 16$). Oznaczmy przez $\mathbf{D} = [d_{ij}] = [d(\gamma_i, \gamma_j)]$ macierz odległości pomiędzy obiektami klasyfikacji. Dla skupienia Γ_p określamy liczbę $\bar{d}_p$ następująco:

$$\bar{d}_p = \frac{1}{n_p(n_p - 1)} \sum_{i=1}^{n_p} \sum_{j=1}^{n_p} d_{ij}^{(p)}, \tag{3}$$

gdzie $d_{ij}^{(p)} = d(\gamma_i, \gamma_j)$, dla $\gamma_i, \gamma_j \in \Gamma_p$ oznacza odległość pomiędzy obiektami należącymi do skupienia Γ_p. Liczba $\bar{d}_p$ jest więc średnią odległością pomiędzy obiektami należącymi do skupienia Γ_p. Jako miary zwartości podziału obiektów przyjęto dwie liczby

$$h_1 = \frac{1}{k} \sum_{p=1}^{k} \bar{d}_p \text{ oraz } h_2 = \max\{\bar{d}_1, \bar{d}_2, ..., \bar{d}_k\}. \tag{4}$$

Z kolei za miarę odrębności podziału przyjęto liczbę H określoną wzorem

$$H = \min_{p,r} \{s_{pr} : p \neq r\}, \tag{5}$$

gdzie $s_{pr} = \min_{i,j} \{d(\gamma_i, \gamma_j) : \gamma_i \in \Gamma_p \text{ i } \gamma_j \in \Gamma_r\}$ jest odległością pomiędzy skupieniami Γ_p i Γ_r równą odległości najbliższych obiektów. Ostatecznie miary jakości klasyfikacji J_1 i J_2 zostały określone następująco:

$$J_1 = h_1/H, J_2 = h_2/H. \tag{6}$$

Im niższe wartości miar J_1 i J_2, tym lepszy podział obiektów. Konstrukcję opisanych miar można znaleźć w pracy [Kolenda 2006].

6. Wyniki

W tab. 2 porównano wyniki klasyfikacji uzyskane za pomocą pięciu metod taksonomicznych. Niektóre metody dały w wyniku te same podziały obiektów. Podziały te oznaczono tymi samymi literami. Można zauważyć m.in., że metoda *k*-średnich daje przeważnie inne podziały obiektów niż metody aglomeracyjne. Metoda średnich połączeń i metoda mediany w pięciu przypadkach podzieliły zbiór obiektów w identyczny sposób. W przypadku sześciu skupień wszystkie metody dały ten sam wynik.

Tabela 2. Porównanie wyników klasyfikacji
(identyczne podziały oznaczono tymi samymi literami)

Metoda	Liczba skupień						
	2	3	4	5	6	7	8
k-średnich	A	B	A	A	A	A	A
Warda	B	B	B	B	A	B	B
Średnich połączeń	C	C	C	C	A	B	C
Mediany	C	D	C	D	A	B	C
Najdalszego sąsiedztwa	B	E	D	C	A	C	B

Źródło: obliczenia własne.

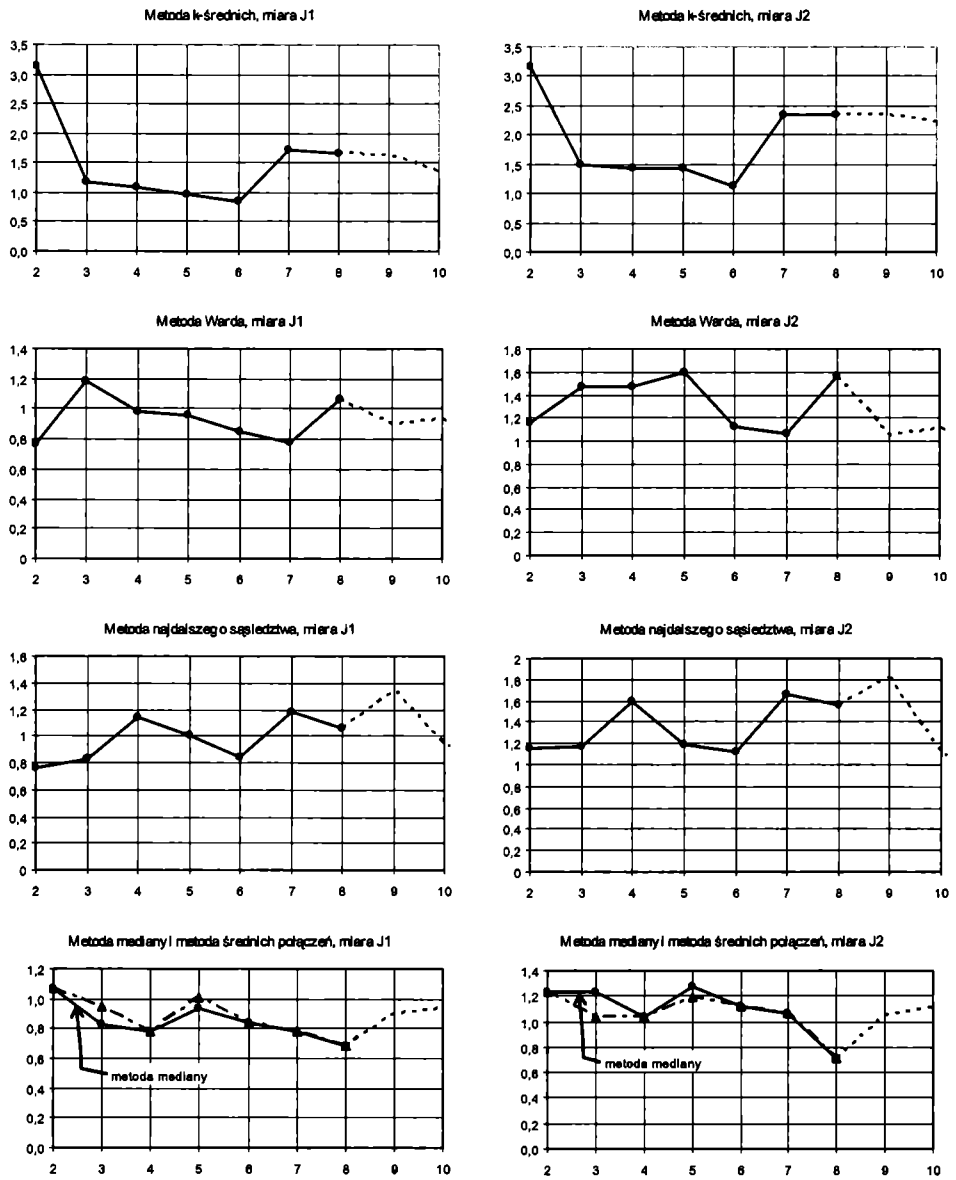
Dla każdego podziału obliczono miary jakości klasyfikacji J_1 i J_2. Na rys. 1 pokazano zmiany wartości wskaźników J_1 i J_2 jako funkcji liczby skupień k. Na wykresach przyjęto szerszy zakres na osi poziomej – od 2 do 10 skupień.

W badanym zakresie (od 2 do 8 skupień) najwyższe wartości mierników odnotowano dla metody *k*-średnich: $J_1 \in [0{,}84; 3{,}15]$, $J_2 \in [1{,}12; 3{,}16]$. Niższe wartości miar otrzymano, stosując metodę najdalszego sąsiedztwa ($J_1 \in [0{,}83; 1{,}18]$, $J_2 \in [1{,}12; 1{,}66]$) oraz metodę Warda ($J_1 \in [0{,}76; 1{,}18]$, $J_2 \in [1{,}06; 1{,}60]$). Najniższe wartości daje metoda mediany ($J_1 \in [0{,}68; 1{,}07]$, $J_2 \in [0{,}71; 1{,}27]$) oraz metoda średnich połączeń ($J_1 \in [0{,}68; 1{,}07]$, $J_2 \in [0{,}71; 1{,}23]$). Pozostając na tym poziomie ogólności, można stwierdzić, że metoda mediany i metoda średnich połączeń dają

Tabela 3. Minimalne wartości miar J_1 i J_2 oraz odpowiadające im liczby skupień

Miara, liczba skupień k	Metoda				
	k-średnich	najdalszego sąsiedztwa	Warda	mediany	średnich połączeń
J_1	0,84	0,83	0,76	0,68	0,68
k	6	2	2	8	8
J_2	1,12	1,12	1,06	0,71	0,71
k	6	6	7	8	8

Źródło: obliczenia własne.



Rys. 1. Zmiany wartości wskaźników jakości J_1 i J_2 jako funkcji liczby skupień (oś pozioma: liczba skupień, oś pionowa: wartość wskaźnika J_1 lub J_2)
Źródło: opracowanie własne.

podziały najlepszej jakości. Minimalne wartości miar J_1 i J_2 dla poszczególnych metod klasyfikacyjnych oraz odpowiadające im liczby skupień zawarto w tab. 3.

7. Wnioski końcowe

1. W przypadku metody *k*-średnich optymalna liczba skupień wynosi 6.

2. Metoda najdalszego sąsiedztwa daje wynik niejednoznaczny – miara J_1 wskazuje na 2 skupienia, a J_2 na 6 skupień. Ponieważ w obu przypadkach wartości wskaźników jakości są zbliżone (dla $k = 2$: $J_1 = 0{,}83$, $J_2 = 1{,}18$; dla $k = 6$: $J_1 = 0{,}85$, $J_2 = 1{,}12$), można w przybliżeniu stwierdzić, że metoda najdalszego sąsiedztwa wskazuje na dwa podziały obiektów: na 2 skupienia i na 6 skupień.

3. Metoda Warda również daje wynik niejednoznaczny – miara J_1 wskazuje na 2 skupienia, a J_2 na 7 skupień. Podobnie jak poprzednio, wartości wskaźników jakości są zbliżone (dla $k = 2$: $J_1 = 0{,}76$, $J_2 = 1{,}15$; dla $k = 7$: $J_1 = 0{,}78$; $J_2 = 1{,}06$) i można przyjąć, że metoda ta również wskazuje na dwa podziały obiektów: na 2 skupienia i na 7 skupień.

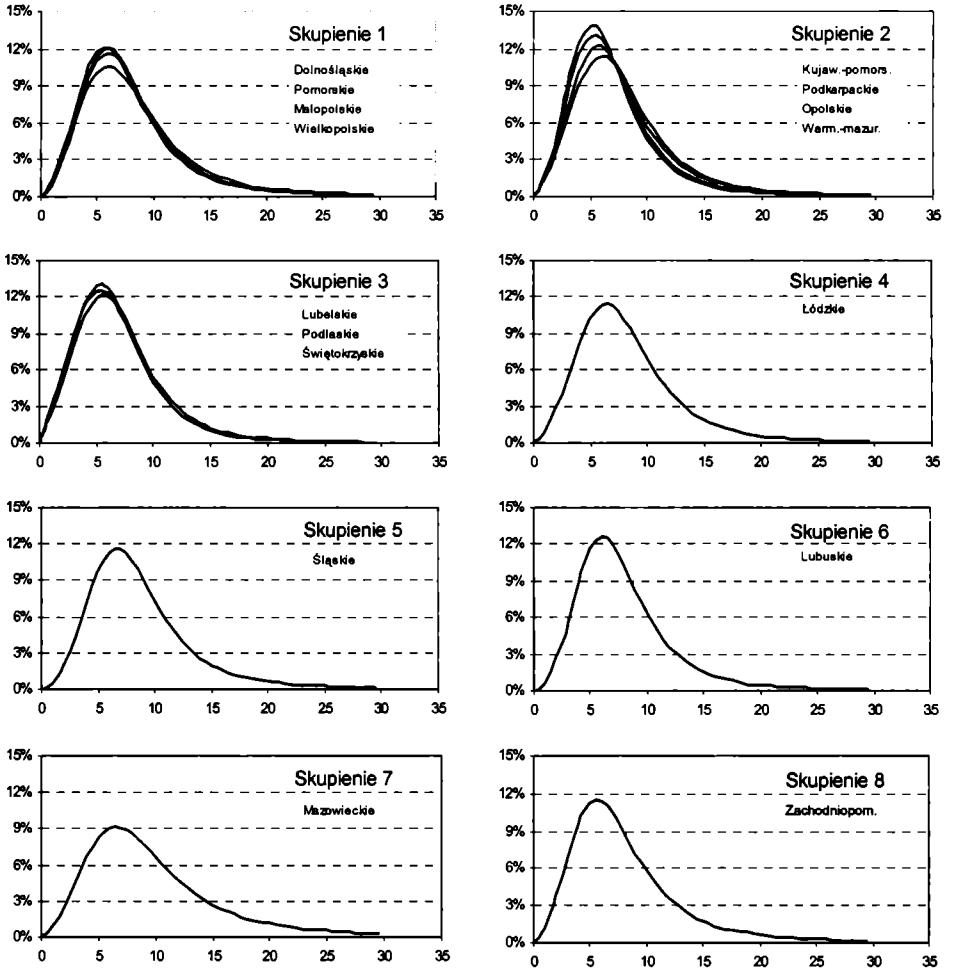
4. Metoda mediany i metoda średnich połączeń dają wynik zgodny i jednoznaczny. W tym przypadku optymalna liczba skupień wynosi 8. Dla tego podziału wartości wskaźników jakości są najmniejsze (wśród wszystkich otrzymanych podziałów).

5. Biorąc po uwagę wszystkie uzyskane podziały, podział na 8 skupień otrzymany metodą mediany (lub metodą średnich połączeń) należy uznać za najlepszy. Podział ten przyjmujemy za ostateczny wynik badania (tab. 4 oraz rys. 2).

Tabela 4. Podział rozkładów dochodów na 8 skupień otrzymany niezależnie metodą mediany oraz metodą średnich połączeń

Województwo	Nr skupienia
Dolnośląskie	1
Małopolskie	1
Pomorskie	1
Wielkopolskie	1
Kujawsko-pomorskie	2
Opolskie	2
Podkarpackie	2
Warmińsko-mazurskie	2
Lubelskie	3
Podlaskie	3
Świętokrzyskie	3
Łódzkie	4
Śląskie	5
Lubuskie	6
Mazowieckie	7
Zachodniopomorskie	8

Źródło: obliczenia własne.



Rys. 2. Podział rozkładów dochodów na 8 skupień otrzymany niezależnie metodą mediany oraz metodą średnich połączeń

Źródło: opracowanie własne.

Literatura

Analiza ekonometryczna kształtowania się płac w Polsce w okresie transformacji (1999), red. S.M. Kot, Wydawnictwo Naukowe PWN, Warszawa-Kraków.

Budżety gospodarstw domowych w 2000 r. (2001), GUS, Warszawa.

Dagum C. (1977), *A New Model of Personal Income Distribution: Specification and Estimation*, „Economie Appliquée" 30 (3), s. 413-437.

Jędrzejczak A. (1990), *Uwagi o zastosowaniu rozkładu Daguma do badania rozkładów płac*, „Wiadomości Statystyczne" nr 7, s. 23-25.
Kolenda M. (2006), *Taksonomia numeryczna, Klasyfikacja, porządkowanie i analiza obiektów wielocechowych*, AE, Wrocław.
Kośny M. (2001), *Probabilistyczne modele rozkładów dochodów – weryfikacja empiryczna*, „Wiadomości Statystyczne" nr 7, s. 20-35.
Łukasiewicz P., Koszela G., Orłowski A. (2007), *Próba klasyfikacji rozkładów dochodów gospodarstw domowych*, „Metody Ilościowe w Badaniach Ekonomicznych" VIII, SGGW, Warszawa, s. 321-329.
Łukasiewicz P., Orłowski A. (2003a), *Rozkłady dochodów jako czynnik zróżnicowania regionalnego*, „Strategie Rozwoju Lokalnego", SGGW, Warszawa, s. 70-80.
Łukasiewicz P., Orłowski A. (2003b), *Ocena sytuacji ekonomicznej gospodarstw wiejskich za pomocą modeli rozkładu dochodów*, „Folia Universitatis Agriculturae Stetinensis" 232, Oeconomica 42, Szczecin, s. 121-127.
Łukasiewicz P., Orłowski A. (2003c), *Probabilistyczne modele rozkładu dochodów polskich gospodarstw domowych*, „Metody Ilościowe w Badaniach Ekonomicznych" III, SGGW, Warszawa, s. 122-130.
Plucińska A., Pluciński E. (2000), *Probabilistyka*, WNT, Warszawa.

HOUSEHOLDS CLASSIFICATION OF INCOME DISTRIBUTIONS

Summary

An income distributions classification of Polish households from 16 provinces of Poland is presented. The resarch material was gained from microobservations issued from households incomes examined by GUS. Empiric income distributions were approximated using Dagum model with three parameters. It enables to identify each of 16 income distributions with certain point in three dimensional space. Objects obtained this way were put under classification using different clustering methods. By the examination of differentiation inside the group and between groups, the best income distributions division was selected.