

Beata Basiura

Akademia Górniczo-Hutnicza w Krakowie

EMPIRYCZNY TEST JEDNORODNOŚCI DLA METODY WARDA STOSOWANEJ DO ANALIZY ZBIORU WOJEWÓDZTW POLSKI

1. Wstęp

Celem pracy jest przedstawienie procedury pozwalającej na obiektywne przerwanie procesu aglomeracji metodą Warda przez odrzucenie hipotezy statystycznej głoszącej, że badana zbiorowość jest jednorodna i nie powinna być dzielona na podzbiory. Podejście to zostało zaproponowane przez Sokołowskiego w 1992 r., a formuły wartości krytycznych podano dla liczby obiektów $w = 49$.

2. Wyznaczenie surowych wartości krytycznych

Kryterium przerywania procesu aglomeracji powinno być oparte na analizie rozkładu prawdopodobieństwa odległości aglomeracyjnej, a więc odległości wskazującej – na każdym etapie łączenia – które obiekty lub grupy obiektów należy połączyć. Ten rozkład szacowano symulacjami komputerowymi, przyjmując następujące założenia:

1. Liczba obiektów w zbiorze jednostek taksonomicznych wynosi $w = 16$ (jest to związane z potrzebą stosowania procedur taksonomicznych w analizach przestrzennego zróżnicowania różnego rodzaju zjawisk w zbiorze województw Polski).
2. Liczba cech k jest liczbą naturalną z przedziału $[2, 20]$.
3. Za model populacji jednorodnej przyjęto wielowymiarowy rozkład normalny o niezależnych składowych.
4. Przyjęto liczbę przebiegów symulacyjnych $n = 100, n = 1000, n = 2000$.

5. Zaproponowana procedura wyznacza wartości krytyczne dla poziomów istotności $\alpha = 0,1$ lub $\alpha = 0,05$.

Wartości krytyczne są wstępnie wyznaczane na podstawie analizy zbiorów wygenerowanych symulacyjnie danych. Wybrane surowe wartości krytyczne odległości aglomeracyjnych zawiera tab. 1. Do generowania prób używano generatora liczb losowych rozkładu normalnego $N(0,1)$, dostępnego w programie Statistica.

Tabela 1. Wybrane surowe wartości krytyczne odległości aglomeracyjnych przy $n = 1000$

w	$\alpha = 0,1$				$\alpha = 0,05$			
	$k = 4$	$k = 9$	$k = 11$	$k = 14$	$k = 4$	$k = 9$	$k = 11$	$k = 14$
16	0,933646	2,300621	2,78143	3,16351	1,016458	2,402860	2,901508	3,280174
15	1,144392	2,585474	3,05575	3,45847	1,208650	2,689176	3,174684	3,565033
14	1,303398	2,794923	3,28808	3,69300	1,393376	2,912878	3,387408	3,790828
13	1,469195	3,014816	3,51050	3,91511	1,544552	3,113371	3,620808	4,045209
12	1,657663	3,226136	3,72736	4,15074	1,757346	3,325757	3,807352	4,238040
11	1,844666	3,468164	3,94926	4,37916	1,940693	3,588560	4,041787	4,482962
10	2,060735	3,712844	4,21310	4,65118	2,178967	3,846105	4,320458	4,783534
9	2,292516	4,000824	4,50405	4,95645	2,419953	4,144415	4,621242	5,060610
8	2,646168	4,348263	4,82329	5,26902	2,778759	4,466070	5,004012	5,395101
7	3,057724	4,714213	5,26814	5,66179	3,200105	4,924500	5,421567	5,831216
6	3,525867	5,243593	5,72309	6,14530	3,679294	5,432462	5,889157	6,321562
5	4,227452	5,850724	6,32040	6,71586	4,427868	6,063607	6,531226	6,933664
4	5,202768	6,701018	7,12905	7,50910	5,464296	6,947684	7,318430	7,751999
3	6,766283	7,948788	8,23240	8,66617	7,113965	8,246283	8,508475	8,956563
2	9,616578	10,151050	10,42095	10,79000	10,319550	10,622120	10,813690	11,246620

Źródło: opracowanie własne.

3. Wyrównanie wartości krytycznych

Wartości krytyczne z tab. 1 wyrównywano za pomocą wybranych funkcji analitycznych dwóch zmiennych $d = d(w, k)$, gdzie w – liczba obiektów, a k – liczba cech. Parametry wszystkich funkcji były istotne statystycznie. Najlepsze dopasowanie wykazała funkcja przybierająca postać:

$$d(w, k) = a * w^{b*k+c} \quad (1)$$

Otrzymano następujące równanie wyrównujące wartości krytyczne dla $\alpha = 0,1$:

$$d(w, k) = 16,18657 * w^{\left(\begin{smallmatrix} 0,02666*k + (-0,93689) \\ (0,000669) \end{smallmatrix} \right)} \quad (2)$$

Współczynnik determinacji w powyższym modelu wynosi $R^2 = 0,979$.

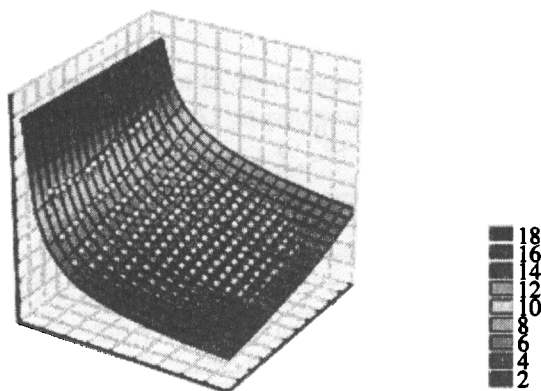
Analogiczne równanie wyrównujące wartości krytyczne dla $\alpha = 0,05$ przybiera postać:

$$d(w, k) = 17,2413 * w^{\left(\begin{smallmatrix} 0,02433*k + (-0,91824) \\ (0,000341) \end{smallmatrix} \right)} \quad (3)$$

Współczynnik determinacji w powyższym modelu wynosi $R^2 = 0,986$.

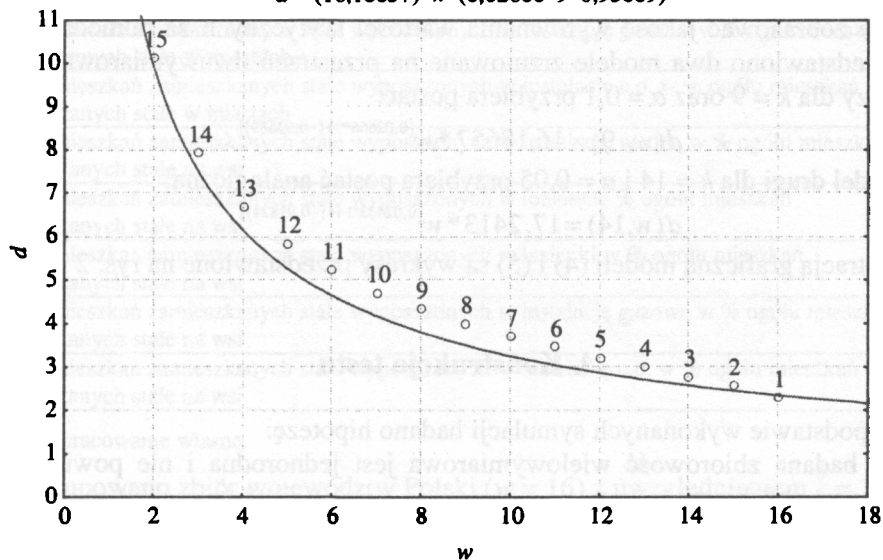
Ilustracją graficzną dopasowania powierzchni do danych symulacyjnych jest rys. 1.

$$\text{Model: } d = a \cdot w^b (b \cdot k + c) \\ d = (17,2413) \cdot w^{(0,02433)} \cdot k + (-0,91824))$$

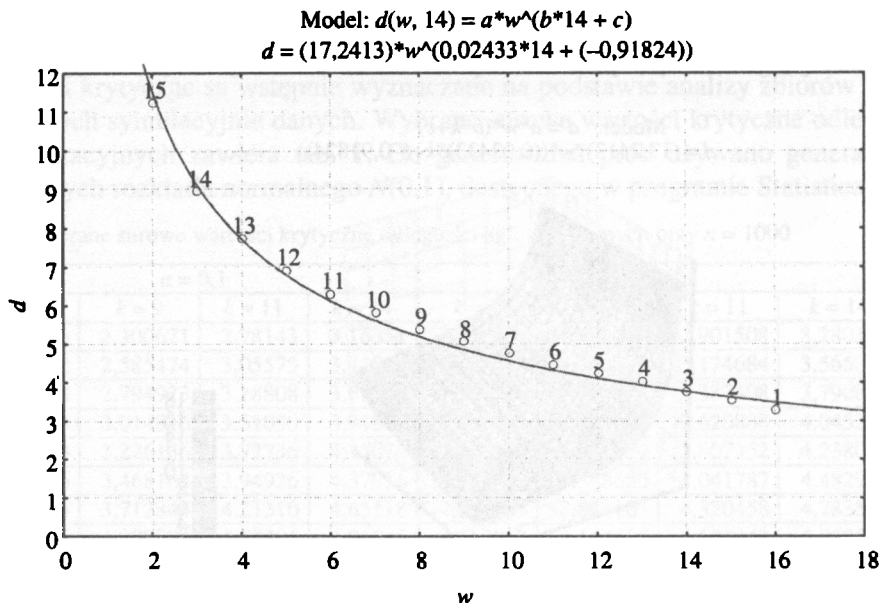


Rys. 1. Wykres dopasowania modelu do danych empirycznych dla $\alpha = 0,05$
Źródło: opracowanie własne.

$$\text{Model: } d(w, 9) = a \cdot w^b (b \cdot 9 + c) \\ d = (16,18657) \cdot w^{(0,02666 \cdot 9 - 0,93689)}$$



Rys. 2. Ilustracja graficzna modelu zrzutowanego przy $k = 9$ i $\alpha = 0,1$
Źródło: opracowanie własne.



Rys. 3. Ilustracja graficzna modelu zrzutowanego przy $k = 14$ i $\alpha = 0,05$
 Źródło: opracowanie własne.

Aby zobrazować jakość wyrównania wartości krytycznych za pomocą funkcji (1) przedstawiono dwa modele zrzutowane na przestrzeń dwuwymiarową. Model pierwszy dla $k = 9$ oraz $\alpha = 0,1$ przybiera postać:

$$d(w, 9) = 16,18657 * w^{(0,02666 * 9 + (-0,93689))} \quad (4)$$

Model drugi dla $k = 14$ i $\alpha = 0,05$ przybiera postać analogiczną:

$$d(w, 14) = 17,2413 * w^{(0,02433 * 14 + (-0,91824))} \quad (5)$$

Ilustracją graficzną modeli (4) i (5) są wykresy przedstawione na rys. 2 i 3.

4. Konstrukcja testu

Na podstawie wykonanych symulacji badano hipotezę:

H_0 : badana zbiorowość wielowymiarowa jest jednorodna i nie powinna być dzielona na podzbiory.

Prawdziwość powyższej hipotezy weryfikowano, porównując odległości z próby d_{obl} z wyrównanymi wartościami krytycznymi $d_{0,1}$ oraz $d_{0,05}$. Proponowany test jest jednostronny z prawostronnym przedziałem krytycznym. Jeżeli na którymś z etapów aglomeracji w kolejnym połączeniu dwóch najbardziej podobnych podgrup trzeba wykorzystać odległość większą od odpowiedniej wartości krytycznej, to proces aglomeracji należy przerwać i jako wynikową uznać klasyfikację otrzymaną

po ostatnim połączeniu. Jeżeli do końca procesu aglomeracji żadna z wykorzystanych odległości nie przekroczy wartości krytycznej, nie ma podstaw do odrzucenia hipotezy o jednorodności.

5. Przykład

Rozważano następujące wskaźniki dotyczące warunków mieszkaniowych w poszczególnych województwach Polski na podstawie danych z Rocznika Statystycznego 2003, zaprezentowane w tab. 2.

Tabela 2. Rozważane wskaźniki warunków mieszkaniowych

Opis
Liczba mieszkań (w tys.) na 1000 ludności
Liczba izb (w tys.) na 1000 ludności
Powierzchnia użytkowa mieszkania w m ² na 1000 mieszkańców
Powierzchnia użytkowa mieszkania w m ² na 1 mieszkańca
Liczba mieszkań zamieszkałych stale wyposażonych w instalację wodną w % ogółu mieszkań zamieszkałych stale w miastach
Liczba mieszkań zamieszkałych stale wyposażonych w toalety w % ogółu mieszkań zamieszkałych stale w miastach
Liczba mieszkań zamieszkałych stale wyposażonych w łazienkę w % ogółu mieszkań zamieszkałych stale w miastach
Liczba mieszkań zamieszkałych stale wyposażonych w instalację gazową w % ogółu mieszkań zamieszkałych stale w miastach
Liczba mieszkań zamieszkałych stale wyposażonych w instalację c.o. w % ogółu mieszkań zamieszkałych stale w miastach
Liczba mieszkań zamieszkałych stale wyposażonych w instalację wodną w % ogółu mieszkań zamieszkałych stale na wsi
Liczba mieszkań zamieszkałych stale wyposażonych w toaletę w % ogółu mieszkań zamieszkałych stale na wsi
Liczba mieszkań zamieszkałych stale wyposażonych w łazienki w % ogółu mieszkań zamieszkałych stale na wsi
Liczba mieszkań zamieszkałych stale wyposażonych w instalację gazową w % ogółu mieszkań zamieszkałych stale na wsi
Liczba mieszkań zamieszkałych stale wyposażonych w instalację c.o. w % ogółu mieszkań zamieszkałych stale na wsi

Źródło: opracowanie własne.

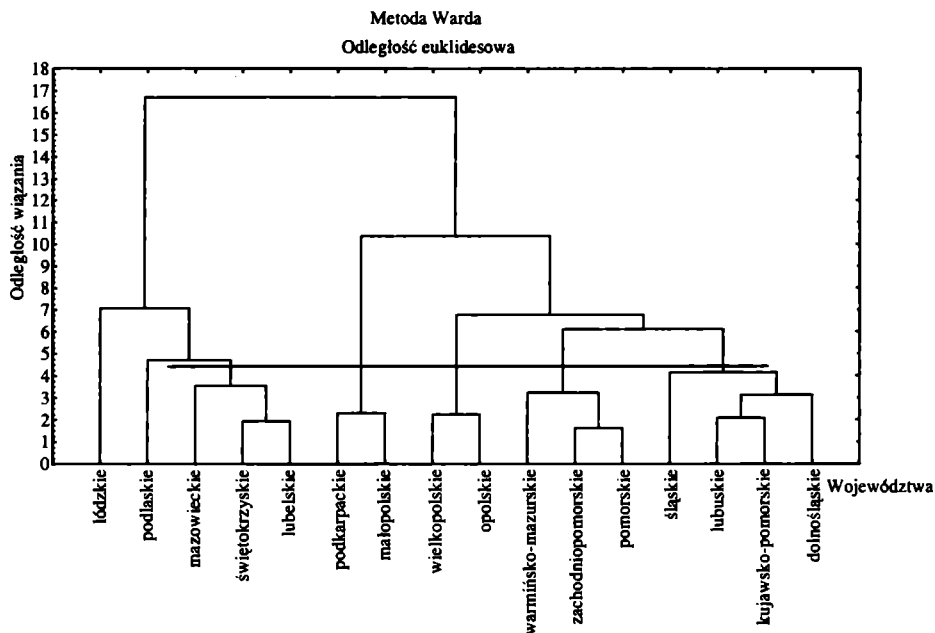
Pogrupowano zbiór województw Polski ($w = 16$) z uwzględnieniem $k = 14$ cech wskaźników mieszkaniowych. Przebieg aglomeracji pokazano na rys. 4.

Porównując w tab. 3 odległości łączenia podgrup z wartościami krytycznymi, okazało się, że odległość aglomeracyjna na poziomie trzech podgrup jest większa od odpowiedniej wartości krytycznej. Zatem na poziomie istotności $\alpha = 0,1$ i $\alpha = 0,05$ hipotezę H_0 należy odrzucić na korzyść tego, że badana zbiorowość nie jest jednorodna i należy ją podzielić na trzy grupy.

Tabela 3. Wartości krytyczne i wartości obliczone odległości aglomeracyjnych

Liczba podgrup	$d_{0,05}$	$d_{0,1}$	d_{obl}
16	3,47573	3,38728	1,614965
15	3,60775	3,51277	1,934601
14	3,75443	3,65206	2,077250
13	3,91863	3,80784	2,253909
12	4,10406	3,98357	2,299270
11	4,31560	4,18381	3,127892
10	4,55984	4,41472	3,236480
9	4,84597	4,68484	3,550974
8	5,18713	5,00641	4,144639
7	5,60305	5,39776	4,723401
6	6,12483	5,88773	6,102473
5	6,80504	6,52497	6,757105
4	7,74120	7,39949	7,079660
3	9,14061	8,70209	10,366930
2	11,55285	10,93648	16,711720

Źródło: opracowanie własne.

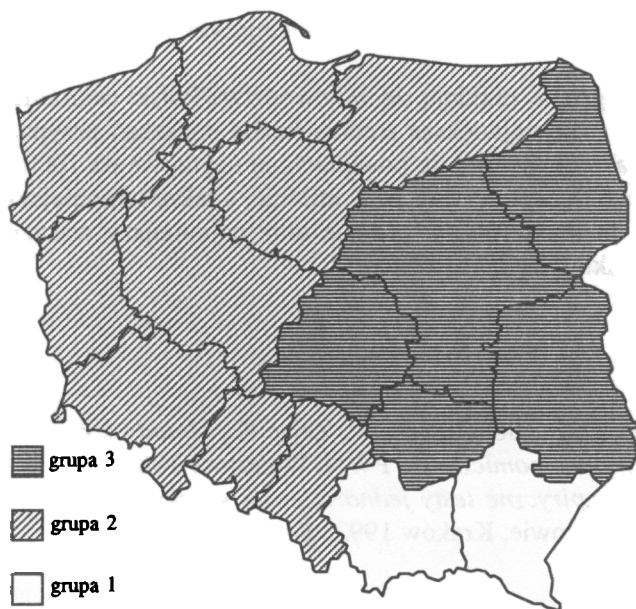


Rys. 4. Dendrogram
Źródło: opracowanie własne.

Tabela 4. Wartości średnie i odchylenia standardowe w grupach

Skrócony opis	Grupa 1		Grupa 2		Grupa 3	
	Mean	StdDev	Mean	StdDev	Mean	StdDev
Liczba mieszkań	270,15	11,1016	304,33	14,192	306,60	18,9136
Liczba izb	1043,45	17,7484	1141,64	36,631	1121,88	48,6183
Pow. użytkowa	20120,50	166,1701	20660,00	1108,807	21162,80	409,6135
Pow. użytkowa/1 mieszk	74,60	2,4042	68,09	4,836	69,28	3,9188
Woda w mieście	98,50	0,4243	99,46	0,219	97,52	0,7050
Toaleta w mieście	96,40	0,7071	95,20	1,792	93,96	2,3158
Łazienka w mieście	95,55	0,3536	92,40	1,796	92,20	3,1432
Gaz w mieście	85,05	5,0205	77,01	4,934	65,18	12,6597
C.o. w mieście	86,45	1,7678	83,39	2,951	87,92	4,2798
Woda na wsi	90,50	2,1213	94,97	1,970	82,04	1,8352
Toaleta na wsi	78,65	3,8891	82,09	3,883	62,76	3,0435
Łazienka na wsi	81,00	3,2527	83,41	4,072	64,78	2,4118
Gaz na wsi	53,50	6,7882	8,43	8,369	7,64	4,8107
C.o. na wsi	64,15	5,1619	68,64	5,591	57,54	5,2204

Źródło: opracowanie własne.



Rys. 5. Mapa Polski z podziałem województw na grupy

Źródło: opracowanie własne.

Wynikiem klasyfikacji był podział województw na następujące grupy:

- grupa 1 – małopolskie i podkarpackie,
- grupa 2 – dolnośląskie, lubuskie, kujawsko-pomorskie, opolskie, pomorskie, śląskie, warmińsko-mazurskie, wielkopolskie, zachodniopomorskie,
- grupa 3 – lubelskie, łódzkie, mazowieckie, podlaskie, świętokrzyskie.

Podział województw na trzy grupy przedstawia rys. 5. W tabeli 4 zawarto wartości średnie poszczególnych cech w wyznaczonych grupach.

6. Wnioski

Grupa 1 to woj. **małopolskie** i **podkarpackie**. Charakterystyczne są dla nich najmniejsza średnia liczba mieszkań i izb przypadających na 1000 mieszkańców, najmniejsza średnia powierzchnia użytkowa na 1000 mieszkańców oraz największa średnia powierzchnia użytkowa przypadająca na 1 mieszkańca. Województwa te mają także największy procentowo dostęp do WC, łazienki i gazu w mieście oraz największy procentowo dostęp do gazu na wsi.

W grupie 2 znalazły się województwa Polski północnej, zachodniej i południowo-zachodniej. Wyróżniają się one największą średnią liczbą izb przypadających na 1000 mieszkańców, największym wskaźnikiem dostępu do wody w mieście oraz największym wskaźnikiem dostępu do wody, WC, łazienki i c.o. na wsi. Te województwa charakteryzują się również najmniejszą średnią powierzchnią użytkową przypadającą na 1 mieszkańca i najmniejszym wskaźnik dostępu do c.o. w mieście.

W ostatniej, 3 grupie znajdują się województwa Polski centralnej i wschodniej. Województwa te mają największą średnią liczbę mieszkań przypadającą na 1000 mieszkańców, największą średnią powierzchnię użytkową na 1000 mieszkańców i największy wskaźnik dostępu do c.o. w mieście, a także najmniejszy wskaźnik dostępu do wody, łazienki, WC w mieście oraz najmniejszy wskaźnik dostępu do wody, WC, łazienki, gazu i c.o. na wsi.

Literatura

- Pociecha J., Podolec B., Sokołowski A., Zajac K., *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa 1988
Sokołowski A., *Empiryczne testy jednorodności w taksonomii*, Zeszyty Naukowe nr 108 AE w Krakowie, Kraków 1992.

EMPIRICAL HOMOGENEITY TEST FOR WARD'S METHOD APPLIED TO THE SET OF POLISH PROVINCES

Summary

A one-sided homogeneity test allowing to stop an agglomerative process in Ward's method is presented in the paper. This approach has been proposed by Sokołowski in 1992, but critical values have been given only for $n = 49$. The re-

sults for $n = 16$ are presented in this paper. Rough critical values are found by analysis of data sets generated by the multidimensional normal distribution. Then they are smoothed using some analytical functions dependent on a number of object and the number of attributes. Functions appeared to fit extremely well. If hypothesis H_0 about the homogeneity is not true then at some stage of sub-groups agglomeration process the agglomerative distances are greater then they should be assuming the hypothesis H_0 is true.