PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS **nr 508 • 2018**

Taksonomia 31
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Mateusz Baryła
Uniwersytet Ekonomiczny w Krakowie
e-mail: mateusz.baryla@uek.krakow.pl

ANALIZA WSKAZAŃ WYBRANYCH MIERNIKÓW SŁUŻĄCYCH OCENIE ZGODNOŚCI DANYCH Z PRAWEM BENFORDA – PRZYPADEK PIERWSZEJ CYFRY ZNACZĄCEJ[1]

ANALYSIS OF INDICATIONS FOR SELECTED MEASURES USED IN ASSESSING DATA CONFORMITY TO BENFORD'S LAW – THE FIRST SIGNIFICANT DIGIT CASE

DOI: 10.15611/pn.2018.508.01
JEL Classification: C10, C38

Streszczenie: Porównywanie empirycznych rozkładów pierwszej cyfry znaczącej, ustalonych na bazie badanych zbiorów danych, z rozkładem Benforda często sprowadza się do konstruowania rankingów tych zbiorów. W tym celu korzysta się z różnych mierników podobieństwa rozkładów, które umożliwiają ocenę stopnia podobieństwa empirycznego rozkładu pierwszej niezerowej cyfry ze wspomnianym rozkładem teoretycznym. W artykule zaprezentowano wyniki analizy wskazań trzynastu mierników podobieństwa rozkładów, opierając się na danych generowanych symulacyjnie. W przeprowadzonym badaniu posłużono się analizą korelacji. Podjęto też próbę klasyfikacji mierników podobieństwa rozkładów za pomocą metody *k*-średnich. Rezultaty grupowania doprowadziły do wniosku, że w badaniach empirycznych, gdy dysponujemy zbiorami danych o różnej wielkości, a celem badań jest stworzenie rankingów analizowanych zbiorów ze względu na stopień ich podobieństwa z rozkładem Benforda, wystarczy ograniczyć się do czterech tego typu mierników.

Słowa kluczowe: mierniki podobieństwa rozkładów, rozkład pierwszej cyfry znaczącej, prawo Benforda, analiza korelacji, analiza skupień.

Summary: Comparing empirical distributions of the first significant digit, which are determined on the basis of the analysed data sets, to Benford's distribution, often leads to the creation of rankings of the data sets. In order to do this, various measures of distributions simi-

[1] Publikacja została dofinansowana ze środków MNiSW przyznanych Wydziałowi Zarządzania Uniwersytetu Ekonomicznego w Krakowie na badania dla młodych naukowców oraz uczestników studiów doktoranckich.

larity are employed, which allow to assess the level of similarity of the first non-zero significant digit empirical distribution to the aforementioned theoretical one. The paper presents the results of indications for thirteen measures of distributions similarity, using simulation data. In the study, the correlation analysis was employed. Moreover, an attempt to classify the measures of distributions similarity was made by means of the ***k***-means method. The outcomes of clustering led to the conclusion that in the empirical research (aimed at the creation of analysed data sets rankings in terms of their level of similarity to Benford's distribution), when one analyses data sets of different sizes, it is enough to use only four measures of this kind.

Keywords: measures of distributions similarity, first significant digit distribution, Benford's law, correlation analysis, cluster analysis.

1. Wstęp

Do grupy metod analizy danych zalicza się m.in. analizę cyfrową (*Digital Analysis*) (pojęcie to zostało wprowadzone w pracy [Nigrini, Mittermaier 1997]), która opiera się na prawie Benforda. Najogólniej rzecz ujmując, prawo Benforda jest wykorzystywane do badania rozkładu cyfr (bądź ich kombinacji) na określonej pozycji znaczącej (bądź pozycjach znaczących) w liczbach. Dlatego bywa ono również nazywane prawem rozkładu cyfr znaczących.

W badaniach empirycznych prawo Benforda znajduje swoje zastosowanie m.in. przy konstruowaniu rankingów analizowanych zbiorów liczbowych ze względu na stopień podobieństwa empirycznego rozkładu pierwszej cyfry znaczącej z rozkładem Benforda (tj. rozkładem wynikającym z prawa Benforda). Rozkład Benforda dla przypadku pierwszej niezerowej cyfry jest określony następującą funkcją prawdopodobieństwa:

$$P(D = d) = \log_{10}(1 + d^{-1}), \tag{1}$$

gdzie: $d \in \{1, 2, ..., 9\}$, a symbol D oznacza zmienną losową opisującą cyfrę występującą na pierwszej pozycji znaczącej w liczbie.

Jako przykład opracowań naukowych, w których tworzono rankingi badanych zbiorów liczbowych, biorąc pod uwagę poziom zgodności empirycznego rozkładu pierwszej cyfry znaczącej z rozkładem Benforda, można wymienić: [Baryła 2017; Rauch i in. 2011; Slijepčević, Blašković 2014]. Podczas konstrukcji tego typu rankingów autorzy posługują się różnymi miernikami, z których jedne są statystykami testowymi (pochodzącymi z testów istotności), inne zaś − miarami, jakie funkcjonują w ramach tzw. statystyki opisowej. Należy w tym miejscu wyraźnie podkreślić, że celem tworzenia takich rankingów jest jedynie uszeregowanie rozpatrywanych zbiorów danych po to, aby sprawdzić, które spośród nich w większym stopniu, a które − w mniejszym stopniu są podobne (niepodobne) do rozkładu Benforda.

Wobec wielości mierników, które dla badaczy mogą stanowić bazę do tworzenia rankingu zbiorów danych, zasadne staje się przeanalizowanie wskazań takich mierników. Autorowi niniejszej pracy znana jest zaledwie jedna publikacja, w której podjęto próbę przeanalizowania wskazań tzw. mierników podobieństwa rozkładów[2].

W pracy [Farbaniec i in. 2012] badano wskazania 17 mierników zgodności rozkładów dla przypadku pierwszej cyfry znaczącej. Podstawą prowadzonego badania było 20 rozkładów empirycznych pierwszej cyfry znaczącej, jakie zawarto w artykule F. Benforda z 1938 r. (zob. [Benford 1938, s. 553]). Rozkłady empiryczne zostały ustalone na podstawie 20 zbiorów danych, których liczebności były zróżnicowane. Prezentując wyniki swoich analiz, autorzy pracy [Farbaniec i in. 2012, s. 166] stwierdzili, że „badania te powinny mieć charakter symulacyjny i opierać się na danych generowanych z kontrolowanym stopniem podobieństwa rozkładów". Wprawdzie stwierdzenie to odnosiło się do problemu oceny diagnostyczności mierników zgodności, niemniej jednak można je potraktować znacznie szerzej.

Celem artykułu jest przeanalizowanie wskazań wybranych mierników zgodności rozkładów (stosowanych do konstrukcji rankingów analizowanych zbiorów danych przy wykorzystaniu rozkładu pierwszej cyfry znaczącej w liczbach), opierając się na danych symulacyjnych. Generowanie empirycznych rozkładów pierwszej niezerowej cyfry było prowadzone z zachowaniem kontroli stopnia niepodobieństwa pomiędzy rozkładem empirycznym a rozkładem Benforda. W badaniu symulacyjnym przeprowadzono analizę korelacji wskazań mierników oraz podjęto próbę ich klasyfikacji.

2. Opis procedury badawczej

Badaniem objęto 13 mierników podobieństwa rozkładów, które − jak się zdaje − są związane z najchętniej wykorzystywanymi przez badaczy metodami[3]. Są to:

- średnia arytmetyczna z dziewięciu statystyk testowych pochodzących z testu dla wskaźnika struktury:

$$\overline{U} = \frac{1}{9}\sum_{d=1}^{9}\sqrt{n}\frac{\left|W_d - p_d^{(B)}\right| - 1/(2n)}{\sqrt{p_d^{(B)}\left(1-p_d^{(B)}\right)}}, \tag{2}$$

- statystyka testowa z testu zgodności chi-kwadrat:

$$\chi^2 = n\sum_{d=1}^{9}\left(W_d - p_d^{(B)}\right)^2 / p_d^{(B)}, \tag{3}$$

- statystyka testowa (3) podzielona przez liczebność próby:

[2] Przez pojęcie miernika podobieństwa (rozkładów) (inaczej: miernika zgodności (rozkładów)) należy rozumieć albo statystykę testową pochodzącą z testu istotności, albo miarę statystyczną funkcjonującą w ramach tzw. statystyki opisowej.

[3] Przykładami opracowań, w jakich wykorzystywano mierniki podobieństwa rozkładów (w ramach metod, z którymi są powiązane), które objęto badaniem w niniejszej pracy, są: [Judge, Schechter 2009; Morzy i in. 2016; Slijepčević, Blašković 2014].

$$\chi^2/n = \sum_{d=1}^{9}\left(W_d - p_d^{(B)}\right)^2/p_d^{(B)}, \tag{4}$$

- statystyka testowa z testu Kołmogorowa-Smirnowa:

$$\lambda = \sqrt{n}\,\sup_d|F_n(d) - F_B(d)|, \tag{5}$$

- statystyka testowa z testu Kuipera:

$$V = \sup_d[F_n(d) - F_B(d)] + \sup_d[F_B(d) - F_n(d)], \tag{6}$$

- statystyka testowa ze zmodyfikowanego testu Kuipera:

$$V^* = V\left(\sqrt{n} + 0{,}155 + 0{,}24n^{-1/2}\right), \tag{7}$$

- miara oparta na odległości Czebyszewa:

$$m = \max_d\left|w_d - p_d^{(B)}\right|, \tag{8}$$

- miara oparta na odległości Euklidesa:

$$\tilde{d}^* = \sqrt{\sum_{d=1}^{9}\left(w_d - p_d^{(B)}\right)^2}\Big/\sqrt{\sum_{d=1}^{8}\left(p_d^{(B)}\right)^2 + \left(1 - p_9^{(B)}\right)^2}, \tag{9}$$

- statystyka testowa z testu bazującego na mierze odległości Czebyszewa:

$$M^* = \sqrt{n}\,\max_d\left|W_d - p_d^{(B)}\right|, \tag{10}$$

- statystyka testowa z testu bazującego na mierze odległości Euklidesa:

$$\widetilde{D}^{**} = \sqrt{n}\sqrt{\sum_{d=1}^{9}\left(W_d - p_d^{(B)}\right)^2}, \tag{11}$$

- statystyka testowa z testu opartego na wartościach przeciętnych porównywanych rozkładów:

$$A^* = |\bar{X}_n - 3{,}4402|/5{,}5598, \tag{12}$$

- średnie bezwzględne odchylenie:

$$MAD = \frac{1}{9}\sum_{d=1}^{9}\left|w_d - p_d^{(B)}\right|, \tag{13}$$

- współczynnik korelacji liniowej pomiędzy porównywanymi rozkładami:

$$r = \left[\sum_{d=1}^{9}\left(w_d - \frac{1}{9}\right)\left(p_d^{(B)} - \frac{1}{9}\right)\right]\Big/\sqrt{\sum_{d=1}^{9}\left(w_d - \frac{1}{9}\right)^2\sum_{d=1}^{9}\left(p_d^{(B)} - \frac{1}{9}\right)^2}. \tag{14}$$

W prezentowanych formułach n oznacza liczebność próby, $p_d^{(B)}$ – prawdopodobieństwo pojawienia się cyfry d zgodnie z rozkładem Benforda ($p_d^{(B)} \equiv P(D = d)$), w_d – częstość względną cyfry d jako pierwszej niezerowej cyfry w zbiorze liczbo-

wym n-elementowym (w statystykach testowych jest ona traktowano jako zmienna losowa, stąd została oznaczona jako W_d), $F_n(d)$ – dystrybuantę empiryczną pierwszej cyfry znaczącej, $F_B(d)$ – dystrybuantę pierwszej niezerowej cyfry w rozkładzie Benforda, $\bar{X}_n$ – średnią arytmetyczną z cyfr występujących na pierwszej pozycji znaczącej w liczbach tworzących n-elementowy zbiór danych.

Aby przeprowadzić badania wskazań mierników zgodności, posłużono się czteroetapową procedurą. Finalnie doprowadziła ona do utworzenia 40 empirycznych rozkładów pierwszej cyfry znaczącej (przy określonych wielkościach prób), niezbędnych do wyznaczenia wartości 13 rozpatrywanych mierników.

Krok pierwszy tej procedury polegał na wygenerowaniu 40 rozkładów teoretycznych pierwszej niezerowej cyfry. Nowo powstałe rozkłady były wyłaniane drogą modyfikacji dokonywanych na rozkładzie Benforda. Poziom niepodobieństwa pomiędzy każdym nowo powstałym rozkładem a rozkładem Benforda był mierzony za pomocą następującego miernika:

$$\delta = \sum_{d=1}^{9} \left| p_d^{(B)} - p_d^{(*)} \right|, \tag{15}$$

przy czym: $p_d^{(B)}$ oznacza prawdopodobieństwo z rozkładu Benforda, a $p_d^{(*)}$ to prawdopodobieństwo z nowego rozkładu teoretycznego.

W badaniu przyjęto, że dla generowanych rozkładów stopień ich niepodobieństwa względem rozkładu Benforda, w sensie miary (15), będzie wahał się w przedziale od 0,005 do 0,20 ze skokiem co 0,005. Łączną liczbę 40 rozkładów teoretycznych otrzymano w wyniku realizacji trzech różnych wariantów.

Wariant I doprowadził do skonstruowania 10 rozkładów (rozkłady $p_d^{(i)}$ dla $i = 1$, 3, 5, ..., 19), które odznaczały się poziomem niepodobieństwa: 0,01 (0,02) 0,19. Dla kolejnych rozkładów spośród możliwych dziewięciu cyfr losowano bez zwracania 2 cyfry, a następnie zwiększano wartości prawdopodobieństw z rozkładu Benforda (odpowiadające wylosowanym cyfrom) o $x/4$, gdzie x to założony poziom niepodobieństwa (za x przyjmowano: 0,01, 0,03, ..., 0,19). Dalej, z pozostałej puli niewylosowanych dotychczas cyfr losowano kolejne dwie (losowanie zależne) po to, by tym razem zmniejszyć wartości prawdopodobieństw z rozkładu Benforda o $x/4$.

Wariant II przyczynił się do utworzenia następnych 10 rozkładów (rozkłady $p_d^{(i)}$ dla $i = 2$, 4, 6, ..., 20), które odznaczały się poziomem niepodobieństwa: 0,02 (0,02) 0,20. Dla kolejnych rozkładów losowano spośród dziewięciu możliwych cyfr 4 z nich (stosując schemat losowania zależnego) i zwiększano prawdopodobieństwa ich wystąpienia z rozkładu Benforda o $x/8$ (tworząc kolejne rozkłady, za x przyjmowano: 0,02, 0,04, ..., 0,20). Dla pozostałych niewylosowanych cyfr pobrano bez zwracania kolejne 4 cyfry, odejmując od skojarzonych z nimi prawdopodobieństw z rozkładu Benforda $x/8$ masy prawdopodobieństwa.

Wariant III owocował otrzymaniem 20 rozkładów (rozkłady $p_d^{(i)}$ dla $i = 21, 22, 23, ..., 40$), które charakteryzowały się poziomem niepodobieństwa: 0,005 (0,01) 0,195. Dla kolejno tworzonych rozkładów losowano liczbę całkowitą a z przedziału [1; 2]. Następnie spośród dziewięciu możliwych cyfr wyłaniano w losowaniu ze zwracaniem a cyfr, dla których od odpowiadających im prawdopodobieństw z rozkładu Benforda odjęto po $x/(2a)$ masy prawdopodobieństwa (przy tworzeniu kolejnych rozkładów w miejsce x podstawiano: 0,005, 0,015, ..., 0,195)[4]. Dalej, ze zbioru wszystkich niewylosowanych dotąd cyfr pobrano (w losowaniu ze zwracaniem) a cyfr, dla których zwiększano prawdopodobieństwa z rozkładu Benforda o $x/(2a)$.

Krok drugi opisywanej procedury polegał na powiązaniu 40 rozkładów teoretycznych z odpowiednio założonymi rozmiarami prób (od $n = 300$ do $n = 117\ 300$ ze skokiem co 3000 obserwacji). Dla kolejno numerowanych ($i = 1, 2, ..., 40$) rozkładów teoretycznych losowano po kolei (bez zwracania) wartości ze zbioru zawierającego liczebności prób {300, 3300, ..., 117 300} i przypisywano je tym rozkładom.

W kroku trzecim utworzono 40 rozkładów empirycznych pierwszej cyfry znaczącej, bazując na otrzymanych rozkładach teoretycznych, powiązanych już z rozmiarami prób. Dla każdego z 40 rozkładów teoretycznych przemnażano prawdopodobieństwa pojawienia się danej cyfry przez skojarzoną z określonym rozkładem liczebność próby. Uzyskane w ten sposób wartości zaokrąglano do liczb całkowitych, a liczby te sumowano, otrzymując tzw. skorygowaną liczebność próby. Skorygowane liczebności prób różniły się od tych podanych w kroku drugim procedury maksymalnie o dwie obserwacje. Opierając się na skorygowanych liczebnościach próby, a także zaokrąglonych wartościach, będących liczbami wystąpień danej cyfry, wyznaczano częstości względne pojawienia się cyfr. Tak otrzymane rozkłady potraktowano jako rozkłady empiryczne.

Ostatni krok charakteryzowanej procedury polegał na obliczeniu wartości 13 mierników zgodności dla uzyskanych w poprzednim kroku 40 rozkładów empirycznych. Otrzymane dane zostały przedstawione w formie tabeli, w której kolejne wiersze były opisywane przez rozkłady empiryczne, a w kolumnach podano wartości mierników. Te dane posłużyły następnie do przeanalizowania wskazań mierników.

3. Rezultaty przeprowadzonego badania

Przed przeprowadzeniem analizy korelacji sprawdzono, czy empiryczne rozkłady mierników podobieństwa podlegają rozkładowi normalnemu. W tym zakresie wykorzystano test Shapiro-Wilka, a rezultaty testowania podano w tab. 1. Jak widać, przy

[4] W opisywanym algorytmie dodatkowo narzucono restrykcję na wyrażenie $x/(2a)$. Gdy wartość tego wyrażenia była większa od mediany obliczonej z wartości prawdopodobieństw pochodzących z rozkładu Benforda (tj. od liczby 0,0792), ponownie losowano liczbę a. Jeżeli okazało się, że w rezultacie odejmowania uzyskano ujemne prawdopodobieństwo dla jakiejś cyfry, to wylosowaną cyfrę odrzucano, a następnie po raz kolejny losowano inną cyfrę do momentu uzyskania wyłącznie dodatnich prawdopodobieństw powstałych w wyniku odejmowania wartości wyrażenia $x/(2a)$.

$\alpha = 0{,}01$ w siedmiu przypadkach na 13 przeprowadzonych testowań należy odrzucić hipotezę zerową zakładającą normalność rozkładu miernika zgodności. Taki wniosek dotyczy następujących mierników: χ^2, χ^2/n, λ, m, M^*, A^*, r.

Tabela 1. Wyniki testowania normalności rozkładów mierników

Miernik zgodności	Wartość statystyki W	p-value	Miernik zgodności	Wartość statystyki W	p-value
$\overline{U}$	0,9459	0,0547	$\tilde{d}^*$	0,9674	0,2976
χ^2	0,7213	0,0000	M^*	0,8645	0,0002
χ^2/n	0,8728	0,0003	$\widetilde{D}^{**}$	0,9324	0,0193
λ	0,9053	0,0027	A^*	0,8943	0,0013
V	0,9328	0,0200	MAD	0,9561	0,1233
V^*	0,9307	0,0170	r	0,8942	0,0013
m	0,9134	0,0048			

Źródło: opracowanie własne.

Tabela 2. Współczynniki korelacji rang Spearmana pomiędzy miernikami podobieństwa rozkładów

	n	$\overline{U}$	χ^2	χ^2/n	λ	V	V^*	m	$\tilde{d}^*$	M^*	$\widetilde{D}^{**}$	A^*	MAD
$\overline{U}$	**0,46**												
χ^2	**0,50**	**0,95**											
χ^2/n	-0,01	**0,73**	**0,77**										
λ	**0,53**	**0,85**	**0,90**	**0,71**									
V	-0,05	**0,67**	**0,70**	**0,92**	**0,74**								
V^*	**0,45**	**0,87**	**0,91**	**0,76**	**0,95**	**0,82**							
m	0,05	**0,59**	**0,72**	**0,90**	**0,74**	**0,88**	**0,75**						
$\tilde{d}^*$	0,00	**0,70**	**0,76**	**0,97**	**0,74**	**0,94**	**0,78**	**0,95**					
M^*	**0,45**	**0,75**	**0,88**	**0,75**	**0,89**	**0,71**	**0,86**	**0,87**	**0,79**				
$\widetilde{D}^{**}$	**0,51**	**0,90**	**0,98**	**0,75**	**0,93**	**0,71**	**0,92**	**0,78**	**0,77**	**0,94**			
A^*	0,09	**0,55**	**0,55**	**0,64**	**0,70**	**0,60**	**0,57**	**0,60**	**0,63**	**0,56**	**0,56**		
MAD	-0,08	**0,75**	**0,70**	**0,91**	**0,65**	**0,90**	**0,72**	**0,75**	**0,91**	**0,60**	**0,67**	**0,59**	
r	0,01	**-0,70**	**-0,76**	**-0,98**	**-0,73**	**-0,93**	**-0,77**	**-0,92**	**-0,98**	**-0,78**	**-0,77**	**-0,62**	**-0,90**

Uwaga: pogrubioną czcionką oznaczono współczynniki korelacji istotne statystycznie na poziomie $\alpha = 0{,}05$.

Źródło: opracowanie własne.

Ponieważ rozkłady większości mierników nie podlegają rozkładowi normalnemu, więc do oceny związków pomiędzy miernikami posłużono się korelacją rangową Spearmana, a nie korelacją liniową. Analiza wyników zawartych w tab. 2 dostarcza trzech podstawowych wniosków. Po pierwsze, do grupy mierników, których wskazania są skorelowane z liczebnością próby, należą: $\overline{U}$, χ^2, λ, V^*, M^*, $\widetilde{D}^{**}$. Po drugie, kierunek uzyskanych korelacji wskazuje, że wszystkie mierniki podobieństwa rozkładów, z wyjątkiem miary r, mają jednakową interpretację, tj. im mniejsze

wartości przyjmują, tym obserwuje się większą zgodność z rozkładem Benforda. W przypadku miernika r sytuacja jest odmienna − im wyższe wartości przyjmuje, tym obserwuje się większą zgodność z rozkładem Benforda. Po trzecie, cztery najsilniejsze korelacje zauważalne są dla następujących par mierników: r i $\tilde{d}^*$, r i χ^2/n, $\widetilde{D}^{**}$ i χ^2, $\tilde{d}^*$ i χ^2/n. Z kolei cztery najsłabsze zależności korelacyjne można dostrzec pomiędzy miernikiem A^* a miernikami: $\overline{U}$, χ^2, M^*, $\widetilde{D}^{**}$.

Macierz danych, jaką uzyskano w wyniku realizacji czteroetapowej procedury opisanej w punkcie drugim, stanowiła też punkt wyjścia analizy taksonomicznej. Każdy z 40 obiektów (rozkładów empirycznych) był opisany przez 13 cech (mierników zgodności). W prowadzonym badaniu podjęto próbę klasyfikacji mierników, które były postrzegane jako punkty rozmieszczone w 40-wymiarowej przestrzeni obiektów. Przy grupowaniu mierników posłużono się cechami znormalizowanymi. Do tego celu wykorzystano metodę unitaryzacji zerowanej (zob. np. [Kukuła 2000]). Wszystkie cechy poza współczynnikiem korelacji liniowej miały charakter destymulant (im wyższą wartość przyjmowały, tym bardziej zmniejszał się poziom podobieństwa do rozkładu Benforda). Natomiast miernik r potraktowano jako stymulantę (im wyższą wartość przyjmował, tym bardziej zwiększał się poziom podobieństwa do rozkładu Benforda).

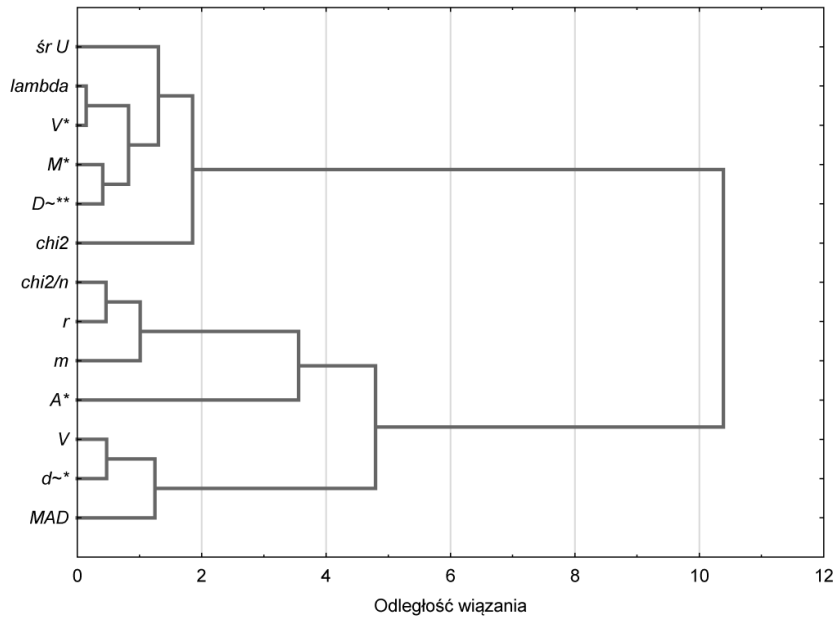
Najpierw wykorzystano metodę Warda (zob. [Ward 1963]) z kwadratową odległością euklidesową po to, aby otrzymane drzewko połączeń stanowiło podstawę do określenia optymalnej liczby podgrup. Wybór metody Warda był podyktowany jej dużą efektywnością w porównaniu do innych metod (zob. np. [Grabiński, Sokołowski 1984]). Ostateczną klasyfikację przeprowadzono z kolei z użyciem metody k-średnich (zob. [Hartigan 1975]), przyjmując tym samym założenie, że skupienia są sferyczne.

Kierując się kryterium pierwszego wyraźnego przyrostu odległości aglomeracyjnej dla kolejnych etapów wiązania (zob. np. [Sokołowski 1992]), otrzymany dendrogram (rys. 1) należy przyciąć na odcinku odległości aglomeracyjnej [1,9; 3,5], co daje 4 podgrupy mierników.

W wyniku realizacji metody k-średnich ($k = 4$) wyróżniono następujące grupy mierników (w nawiasach podano odległości od środka właściwego skupienia):

- grupa 1: $\overline{U}$ (0,1142), λ (0,0602), V^* (0,0580), M^* (0,1054), $\widetilde{D}^{**}$ (0,0490),
- grupa 2: χ^2/n (0,1028), m (0,1008), A^* (0,1827), r (0,0941),
- grupa 3: χ^2 (0,00),
- grupa 4: V (0,0805), $\tilde{d}^*$ (0,0682), MAD (0,1020).

Uzyskane wyniki badania symulacyjnego pokazują, że w analizach empirycznych, ukierunkowanych na konstruowanie rankingów rozpatrywanych zbiorów liczbowych ze względu na poziom podobieństwa empirycznego rozkładu pierwszej cyfry znaczącej z rozkładem Benforda, nie ma konieczności wykorzystywania wszystkich 13 mierników. Dysponując zbiorami danych o zróżnicowanej wielkości, wystarczy uwzględnić jedynie 4 mierniki zgodności: miernik χ^2 oraz po jednym reprezentancie z grup oznaczonych numerami: 1, 2 i 4.



Rys. 1. Dendrogram dla 13 mierników podobieństwa rozkładów

Źródło: opracowanie własne.

4. Zakończenie

Pomimo licznych założeń, jakie poczyniono w prowadzonym badaniu, otrzymane wyniki stanowią cenną wskazówkę przy prowadzeniu analiz empirycznych, mających na celu porównywanie (z wykorzystaniem mierników podobieństwa rozkładów) badanych zbiorów liczbowych (pod kątem poziomu zgodności empirycznych rozkładów pierwszej niezerowej cyfry liczb z tych zbiorów z rozkładem Benforda) w sytuacji, gdy liczebności takich zbiorów są zróżnicowane. Ciekawym zagadnieniem wydaje się przeanalizowanie wskazań mierników zgodności nie tylko przy założeniu zmiennej liczebności próby, lecz i w wariancie jej stałej liczebności. Co więcej, badanie tego typu można dodatkowo rozszerzyć o inne przypadki analiz cyfrowych (np. przypadek drugiej cyfry znaczącej czy dwóch pierwszych cyfr znaczących).

Literatura

Baryła M., 2017, *Analiza rozkładu pierwszej cyfry znaczącej danych finansowych wybranych spółek z sektora mediów notowanych na GPW w Warszawie*, [w:] Jajuga K., Walesiak M. (red.), *Klasyfikacja i analiza danych − teoria i zastosowania*, Taksonomia 29, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 469, s. 11-20.

Benford F., 1938, *The Law of Anomalous Numbers*, Proceedings of the American Philosophical Society, vol. 78, no. 4, s. 551-572.

Farbaniec M., Grabiński T., Zabłocki B., Zając W., 2012, *Metody oceny zgodności rozkładów cyfr znaczących z prawami Benforda*, [w:] Chmielowski W.Z., Wilk-Kołodziejczyk D. (red.), *Metody analizy i oceny bezpieczeństwa oraz jakości informacji*, Krakowskie Towarzystwo Edukacyjne sp. z o.o. Oficyna Wydawnicza AFM, Kraków, s. 143-178.

Grabiński T., Sokołowski A., 1984, *Z badań nad efektywnością wybranych procedur taksonomicznych*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, nr 181, s. 63-80.

Hartigan J.A., 1975, *Clustering Algorithms*, John Wiley & Sons, New York.

Judge G., Schechter L., 2009, *Detecting Problems in Survey Data Using Benford's Law*, The Journal of Human Resources, vol. 44, no. 1, s. 1-24.

Kukuła K., 2000, *Metoda unitaryzacji zerowanej*, Wydawnictwo Naukowe PWN, Warszawa.

Morzy M., Kajdanowicz T., Szymański B.K., 2016, *Benford's Distribution in Complex Networks*, Scientific Reports, 6:34917, DOI: 10.1038/srep34917.

Nigrini M.J., Mittermaier L.J., 1997, *The Use of Benford's Law as an Aid in Analytical Procedures*, Auditing: A Journal of Practice & Theory, vol. 16, no. 2, s. 52-67.

Rauch B., Göttsche M., Brähler G., Engel S., 2011, *Fact and Fiction in EU-Governmental Economic Data*, German Economic Review, vol. 12, issue 3, s. 243-255.

Slijepčević S., Blašković B., 2014, *Statistical detection of fraud in the reporting of Croatian public companies*, Financial Theory and Practice, vol. 38, no. 1, s. 81-96.

Sokołowski A., 1992, *Empiryczne testy istotności w taksonomii*, Zeszyty Naukowe, Seria specjalna: Monografie, nr 108, Akademia Ekonomiczna w Krakowie, Kraków.

Ward J.H., 1963, *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, vol. 58, issue 301, s. 236-244.