

| Tadeusz Grabiński

# Uproszczona metoda delimitacji wektorowej

## Streszczenie

W artykule przedstawiono propozycję graficznej procedury porządkowania i grupowania obiektów opisanych pojedynczą cechą ujętą w formie szeregu przekrojowego, czasowego lub przekrojowo-czasowego. Metoda ta nawiązuje do klasycznej metody porządkowania opracowanej ponad 100 lat temu przez Jana Czekanowskiego. Zaletami proponowanej procedury są mała pracochłonność, intuicyjność i możliwość realizacji za pomocą arkusza kalkulacyjnego z zastosowaniem kilku elementarnych funkcji. Metodę wykorzystano do analizy taksonomicznej zbioru 28 krajów Unii Europejskiej opisanych trzema cechami: stopa bezrobocia, zagrożenie ubóstwem oraz długość życia, odnoszącymi się do dwóch lat – 2010 oraz 2015 r.

**Słowa kluczowe:** taksonometria, porządkowanie, grupowanie, delimitacja, szeregi przekrojowo-czasowe.

**Klasyfikacja JEL:** C38.

## 1. Wprowadzenie

W artykule przedstawiono propozycję procedury delimitacji (porządkowania i grupowania) obiektów wymagającej znajomości kilku podstawowych funkcji arkusza kalkulacyjnego oraz niewielkiego nakładu czasu pracy. Jej dodatkową zaletą jest intuicyjność i możliwość śledzenia rezultatów we wszystkich etapach.

Podstawą analizy jest wektor danych źródłowych wyrażonych w tych samych jednostkach pomiarowych (strumieniowych lub wskaźnikowych). Może to być

szereg przekrojowy (zbiór obiektów opisanych jedną cechą w jednym okresie), szereg czasowy (pojedynczy obiekt opisany jedną cechą w kilku okresach) lub szereg przekrojowo-czasowy (zbiór obiektów opisanych jedną cechą w kilku okresach).

Nie jest to więc metoda *stricte* z obszaru taksonometrii wielowymiarowej, gdzie analizie podlega zbiór cech. Tym niemniej podstawą delimitacji w prezentowanej metodzie może być także cecha syntetyczna uzyskana z agregacji kilku cech cząstkowych metodami wielowymiarowej analizy porównawczej. Dlatego też delimitacja oparta na wektorze danych również może mieć charakter wielowymiarowy. Innym sposobem uzyskania wielowymiarowości delimitacji wektorowej jest przeprowadzenie analizy dla kilku cech i porównanie uzyskanych wyników.

## 2. Charakterystyka danych źródłowych

Omawianą metodę delimitacji wektorowej wykorzystano do analizy taksonometrycznej zbioru 28 krajów Unii Europejskiej opisanych trzema cechami: stopa bezrobocia (B), zagrożenie ubóstwem (U) oraz długość życia (D), odnoszącymi się do dwóch lat (2010 i 2015).

Dane zaczerpnięto z *Rocznika statystyki międzynarodowej 2016* (<http://stat.gov.pl/obszary-tematyczne/roczniki-statystyczne/roczniki-statystyczne/rocznik-statystyki-miedzynarodowej-2016,10,5.html>). W tabeli 1 podano wykaz krajów podlegających analizie oraz przyjęte dla nich skróty wykorzystywane w diagramach.

Tabela 1. Wykaz analizowanych krajów

N	Kraj	Skrót	N	Kraj	Skrót	N	Kraj	Skrót	N	Kraj	Skrót
1	Austria	AT	8	Estonia	EE	15	Irlandia	IE	22	Polska	PL
2	Belgia	BE	9	Grecja	EL	16	Włochy	IT	23	Portugalia	PT
3	Bułgaria	BG	10	Hiszpania	ES	17	Litwa	LT	24	Rumunia	RO
4	Cypr	CY	11	Finlandia	FI	18	Luksemburg	LU	25	Szwecja	SE
5	Czechy	CZ	12	Francja	FR	19	Łotwa	LV	26	Słowenia	SI
6	Niemcy	DE	13	Chorwacja	HR	20	Malta	MT	27	Słowacja	SK
7	Dania	DK	14	Węgry	HU	21	Holandia	NL	28	Wielka Brytania	UK

Źródło: opracowanie własne.

Na podstawie tych informacji wyznaczono następujące cechy wynikowe:

1) statyczne wartości standaryzowane ustalane odrębnie dla każdego roku i każdej cechy z wykorzystaniem funkcji Excela *procent.pozycja*, zgodnie z którą

maksymalnej wartości w zbiorze podlegającym standaryzacji przypisuje się 1 (100%), natomiast wartości najmniejszej – 0 (0%)<sup>1</sup>. W przypadku długości życia standaryzacji dokonywano według zmienionej formuły [1-procent.pozycja], dzięki czemu wszystkie 3 cechy uzyskały jednakowy status, wyrażający się w ujednoliconej ocenie ich wielkości, zgodnie z regułą: im niższe są wartości każdej z tych cech, tym korzystniej należy ocenić dany kraj;

2) dynamiczne wartości standaryzowane wyznaczane według analogicznych zasad jak wartości statyczne, z tą różnicą, że standaryzacji nie dokonuje się odrębnie dla każdego roku, lecz dla połączonych wartości danej cechy z lat 2010 i 2015. Oznacza to, że wartości krańcowe (0–1) przypisywane są krajom z najbardziej i najmniej korzystną wartością danej cechy spośród jej wartości z tych lat;

3) wartości zagregowane zdefiniowane jako średnie arytmetyczne z wartości standaryzowanych z obydwu lat w ujęciu statycznym [%B], [%U], [%D] oraz dynamicznym [%%B], [%%U], [%%D];

4) statyczne wskaźniki syntetyczne jako średnie arytmetyczne ze statycznych wartości standaryzowanych dla wszystkich cech B-U-D w trzech wariantach: [%X10] – dla 2010 r., [%X15] – dla 2015 r. oraz [%X] – dla obydwu lat;

5) dynamiczne wskaźniki syntetyczne jako średnie arytmetyczne z dynamicznych wartości standaryzowanych dla wszystkich cech B-U-D, także w trzech wariantach: [%%X10] – dla 2010 r., [%%X15] – dla 2015 r. oraz [%%X] – dla obydwu lat.

Tabela 2. Podstawowe parametry opisowe zmiennych wyjściowych

Parametr	Stopa bezrobocia			Zagrożenie ubóstwem			Długość życia		
	B10	B15	B	U10	U15	U	D10	D15	D
Min	4,4	4,9	5,2	14,4	14,0	14,2	70,7	74,5	72,7
Q1	6,3	6,8	6,6	18,8	18,4	18,6	75,1	77,7	76,3
Q2	7,7	8,9	8,8	22,5	23,5	23,2	79,6	81,3	80,5
Q3	10,1	11,5	11,2	27,7	28,8	27,9	80,0	81,9	81,0
Max	16,4	26,3	20,3	49,2	41,3	45,3	81,6	83,3	82,4
Średnie arytmetyczne	8,5	10,4	9,5	24,6	24,3	24,4	77,7	79,9	78,8
Odchylenia standardowe	3,05	5,22	3,67	8,35	6,96	7,50	3,36	2,91	3,11
V1 = odch./śred. (w %)	36	50	39	34	29	31	4	4	4
V2 = (Q3–Q1)/Q2 (w %)	49	53	53	40	44	40	6	5	6

Źródło: opracowanie własne.

<sup>1</sup> W nowszych wersjach programu Excel (2013, 2016) zamiast funkcji *procent.pozycja* należy skorzystać z funkcji *proc.poz.przedz.otw* lub *proc.poz.przedz.zamk*.

Liczbę zmiennych można jeszcze powiększyć, np. definiując zmienne syntetyczne o zróżnicowanych wagach lub sięgając po inne metody standaryzacji bądź poszerzając liczbę lat. W tabeli 2 przytoczono podstawowe parametry opisowe wyjściowych zmiennych (kwartyle, średnie arytmetyczne, odchylenia standardowe, współczynniki zmienności). Warto tu zwrócić uwagę na niewielkie (4–6%) zróżnicowanie krajów pod względem długości życia w odróżnieniu od pozostałych dwóch zmiennych, dla których współczynniki zmienności zawierają się w przedziale 30–50%.

### 3. Algorytm delimitacji wektorowej

Proponowana procedura delimitacji nawiązuje do najstarszej, liczącej ponad 100 lat, taksonometrycznej metody Czekanowskiego (por. [Czekanowski 1913, Piasecki 1971]). Wspólną cechą obydwu metod jest wykorzystanie diagraficznego macierzy odległości, na podstawie której można ustalić, czy istnieją grupy podobnych obiektów, a jeżeli tak, to czy możliwa jest identyfikacja tych grup. Prezentowana metoda różni się od metody Czekanowskiego procedurą porządkowania macierzy odległości. Zamiast dokonywać sekwencyjnych przestawień wierszy i kolumn macierzy odległości, szukając optymalnej wartości funkcji kryterium określającej stopień uporządkowania wartości w macierzy<sup>2</sup>, przyjmuje się, że porządek wierszy i kolumn w macierzy odległości wynika z wybranej zmiennej delimitacyjnej.

Grupy obiektów ustala się, zaokrąglając wartości zmiennej delimitacyjnej do coraz to większych wartości. Drugim możliwym rozwiązaniem jest zaokrąglanie elementów macierzy odległości wyznaczonej na podstawie zmiennej delimitacyjnej. W utworzonej aplikacji przewidziano wykorzystanie obydwu możliwości zaokrągleń, zarówno wyjściowych wartości zmiennej delimitacyjnej, jak i wyznaczonych na jej podstawie elementów macierzy odległości.

Poniżej przedstawiono kolejne kroki w proponowanej metodzie.

1. Punktem wyjścia analizy jest wektor wartości dowolnej zmiennej delimitacyjnej. Może to być w szczególności zmienna syntetyczna reprezentująca zbiór innych zmiennych. Może to być zmienna w postaci szeregu czasowego. W tym przypadku uzyskuje się delimitację okresów, co określane jest jako periodyzacja. Wyjściowa zmienna nie może mieć zerowej wariancji. Ponadto powinna być wyrażona w jednostkach pozwalających na wykonywanie podstawowych operacji algebraicznych, np. wyznaczanie różnic lub ilorazów wartości.

2. Wartości zmiennej delimitacyjnej standaryzuje się za pomocą formuły oddającej relacje strukturalne wartości oryginalnych. Zakładając, że obliczenia wykonywane są w arkuszu kalkulacyjnym, proponuje się w tym celu wykorzy-

---

<sup>2</sup> Przykłady takich algorytmów można znaleźć w pracach: [Kozłowski 1972, Sołtysiak 1977a, b].

stać funkcję *procent.pozycja*, a w nowszych wersjach Excela funkcję *proc.poz.przedz.zamk*.

3. Wartości standaryzowanej zmiennej delimitacyjnej porządkuje się w kolejności rosnącej lub malejącej. Występuje tu problem identycznych wartości zmiennej, których pojawianie się nie pozwala wykonać niektórych operacji numerycznych przeprowadzanych w proponowanej metodzie. Dlatego też dokonuje się modyfikacji wszystkich wartości zmiennej zgodnie ze wzorem:

$$x'_i = x_i + x_i \cdot \text{los}()/1000 \quad (i = 1, 2, \dots, n), \quad (1)$$

gdzie *los()* to generator liczb losowych z przedziału (0, 1). Dzięki tej operacji w zbiorze wartości zmiennej delimitacyjnej nie ma identycznych realizacji.

4. Generuje się kwadratową macierz odległości o wymiarach  $n \times n$ , w której umieszcza się moduły różnic pomiędzy wszystkimi wartościami uporządkowanego wektora realizacji zmiennej delimitacyjnej:

$$d_{ij} = |x'_i - x'_j| \quad (i, j = 1, \dots, n). \quad (2)$$

5. Elementy macierzy modułów różnic formatuje się za pomocą funkcji Excela, zastępując wartości (2) symbolami graficznymi identyfikującymi rząd ich wielkości (formatowane warunkowe). W wyniku tej operacji na głównej przekątnej macierzy odległości pojawiają się symbole graficzne przypisane najniższym wartościom różnic (na głównej przekątnej będą to wartości zerowe). Im bardziej oddalone od głównej przekątnej są komórki macierzy (2), tym większe są tam wartości, co jest uwidocznione za pomocą ikon graficznych w różnych kolorach lub kształtach.

6. Grupy podobnych obiektów tworzą kwadraty składające się z ikon przypisanych najmniejszym wartościom modułów różnic. Jeżeli wynikowa macierz nie pozwala zidentyfikować wyraźnych ugrupowań (kwadratów identycznych ikon), to przeprowadza się procedurę stopniowego zaokrąglania wartości zmiennej delimitacyjnej (1). Do tego celu służą funkcje arkusza kalkulacyjnego Excel: *zaokr.w.góre*, *zaokr.w.dół*.

7. W analizach wykorzystano funkcje *zaokr.w.dół* z parametrem przyjmującym kolejno wartości: 0,1%, 1%, 2%, 3%, 4% itd. Pierwszy parametr [0,1%] generuje macierz podobieństwa na podstawie oryginalnych wartości zmiennej delimitacyjnej zaokrąglonych do 1 promila. Drugi parametr [1%] zaokrągla wartości zmiennej do 1%, kolejny parametr [2%] – do wielokrotności 2% itd. Na przykład przy zastosowaniu funkcji *zaokr.w.dół* z parametrem [2%] wektor wartości [0,13; 0,14; 0,17; 0,20] przyjmuje postać [0,12; 0,14; 0,16; 0,20], natomiast z parametrem [3%] wektor wynikowy ma składowe [0,12; 0,12; 0,15; 0,18].

8. Drugim narzędziem poszukiwania rozłącznych ugrupowań obiektów jest zaokrąglanie elementów macierzy odległości (2) za pomocą tej samej funkcji

*zaokr.w.dół.* W praktyce dobrze jest rozpocząć analizę od wartości oryginalnych, przyjmując obydwa parametry zaokrągleń (dla wartości zmiennej delimitacyjnej oraz dla elementów macierzy odległości) na poziomie 0,1%. Następnie należy stopniowo zmieniać ten parametr od 1% do 5–7%, najpierw dla elementów macierzy odległości, a następnie dla zmiennej delimitacyjnej.

9. Wprowadzanie do arkusza kolejnych parametrów określających rząd zaokrągleń generuje macierze o innej konfiguracji ikon. Wybór finalnej konfiguracji, a tym samym rezultat delimitacji proponuje się ustalać z zastosowaniem następującej procedury:

– na podstawie uporządkowanych rosnąco zaokrąglonych wartości zmiennej delimitacyjnej wyznacza się ciąg różnic pomiędzy sąsiednimi jej wartościami:

$$g_i = x'_i - x'_{i-1} \quad (i = 2, 3, \dots, n). \quad (3)$$

W odrębnej kolumnie zapisuje się niezerowe wartości różnic  $g_i$ , eliminując ich wartości zerowe;

– ustala się minimalną wielkość różnic  $g_i$  (po wyeliminowaniu wartości zerowych), a następnie w kolejnej kolumnie zapisuje wartości tych różnic większe od minimalnej;

– operacje z poprzedniego punktu powtarza się aż do momentu, w którym wszystkie różnice  $g_i$  zostaną wyeliminowane;

– wiersze, w których pojawi się najwięcej wartości różnic, wyznaczają delimitację zbioru obiektów. Są to miejsca, w których obserwuje się największe różnice pomiędzy sąsiednimi elementami uporządkowanych wartości zmiennej delimitacyjnej.

10. Druga metoda ustalania liczby ugrupowań opiera się na sumach komórek zawierających wartości zaliczone do kategorii wskazującej na maksymalne podobieństwo obiektów, wyznaczonych dla każdego wiersza. Miejsca podziału obiektów na grupy są zlokalizowane w wierszach, w których liczba maksymalnych podobieństw jest najmniejsza. Podobnie jak w poprzedniej metodzie można tu sekwencyjnie eliminować maksymalne liczby maksymalnych podobieństw, dochodząc do momentu, w którym pozostaną tylko wielkości minimalne.

#### 4. Wyniki delimitacji krajów Europy

Przedstawioną procedurę delimitacji wykorzystano do grupowania krajów Europy, wykorzystując omówione dane źródłowe. Wyniki analizy zebrano w tabelach 3–10, zawierających diagramy dla:

– wyjściowych wartości zmiennej delimitacyjnej bez zaokrągleń, ze standaryzacją dynamiczną (tabela 3) oraz statyczną (tabela 4),

– zaokrąglonych wartości zmiennej delimitacyjnej na poziomie 5%, ze standaryzacją dynamiczną (tabela 5) oraz statyczną (tabela 6),

- zaokrąglonych wartości zmiennej delimitacyjnej na poziomie 20%, ze standaryzacją dynamiczną (tabela 7) oraz statyczną (tabela 8),
- zaokrąglonych wartości zmiennej delimitacyjnej na poziomie 30%, ze standaryzacją dynamiczną (tabela 9) oraz statyczną (tabela 10).

W każdym diagramie pogrubionymi krawędziami oznaczono grupy krajów o podobnych wartościach zmiennej delimitacyjnej. Podział krajów na grupy przeprowadzany jest tak, aby w kwadratach wzdłuż głównej przekątnej znajdowało się jak najwięcej komórek z identyfikatorami wskazującymi na maksymalne podobieństwo krajów. Elementom macierzy odległości przypisano różne kolory. Wartościom mierników wskazujących na najmniejszy poziom podobieństw klasyfikowanych obiektów przyporządkowany jest kolor zielony, kolor pomarańczowy wskazuje na przeciętny poziom podobieństwa obiektów, a kolor czerwony na poziom wysoki. Mierniki odległości pomiędzy obiektami o największym stopniu podobieństwa oznaczone są kolorem czarnym. Z definicji są to elementy głównej przekątnej macierzy odległości wskazujące na podobieństwo obiektu do samego siebie. Konfiguracja geometryczna elementów macierzy oznaczonych czarnym kolorem wskazuje na stopień separowalności analizowanego zbioru obiektów. Konfiguracje w formie dużych kwadratów pozwalają podzielić zbiór obiektów na odpowiednią liczbę rozłącznych, wyraźnie separowalnych grup. Natomiast konfiguracja w postaci wstęgi wzdłuż głównej przekątnej macierzy odległości oznacza sytuację, w której trudno jest wydzielić grupy obiektów podobnych do siebie i jednocześnie niepodobnych do obiektów zaliczonych do innych grup.

Przy ustalaniu grup krajów pomocne są zamieszczone po prawej stronie diagramów:

- maksymalne wartości różnic zmiennej delimitacyjnej,
- liczby komórek zawierających identyfikatory maksymalnego podobieństwa.

Ponadto dla celów porównawczych dokonano delimitacji krajów Europy klasyczną metodą Czekanowskiego. Wykorzystano w tym celu program do grupowania danych MaCzek, wersja 3.3.44, udostępniony na stronie <http://www.antropologia.uw.edu.pl/MaCzek>. Podstawą analizy w tym przypadku były wektory zawierające statyczne (%X) oraz dynamiczne (%%X) wskaźniki syntetyczne wyznaczone na podstawie danych z obydwu lat. Wyniki delimitacji zebrano w tabelach 11 i 12. Znajdują się tu macierze Czekanowskiego dla danych:

- nieuporządkowanych (według alfabetycznej kolejności krajów) – tabela 11, cecha [%%X],
- uporządkowanych za pomocą algorytmu uproszczonego – tabela 11, cecha [%%X],
- uporządkowanych za pomocą algorytmu genetycznego – tabela 12, cecha [%X],
- uporządkowanych za pomocą algorytmu UMCzek – tabela 12, cecha [%X].

Tabela 3. Delimitacja krajów Europy na podstawie zmiennej %%X bez zaokrągleń

Table with 28 rows and columns for countries (NL, AT, LU, SE, CZ, DK, FR, MT, FI, SI, UK, BE, IT, CY, IE, ES, PL, EE, SK, EL, RO, HU, HR, BG, LT, LV) and numerical values for %%X. Includes columns for indices 2-5 and 0.1.

Źródło: opracowanie własne.

Tabela 4. Delimitacja krajów Europy na podstawie zmiennej %X bez zaokrągleń

Table with 28 rows and columns for countries (NL, AT, LU, SE, CZ, FR, DE, DK, FI, SI, UK, BE, IT, CY, ES, IE, PL, EE, PT, EL, HU, RO, HR, BG, LT, LV) and numerical values for %X. Includes columns for indices 1, 5, and cz.

Źródło: opracowanie własne.



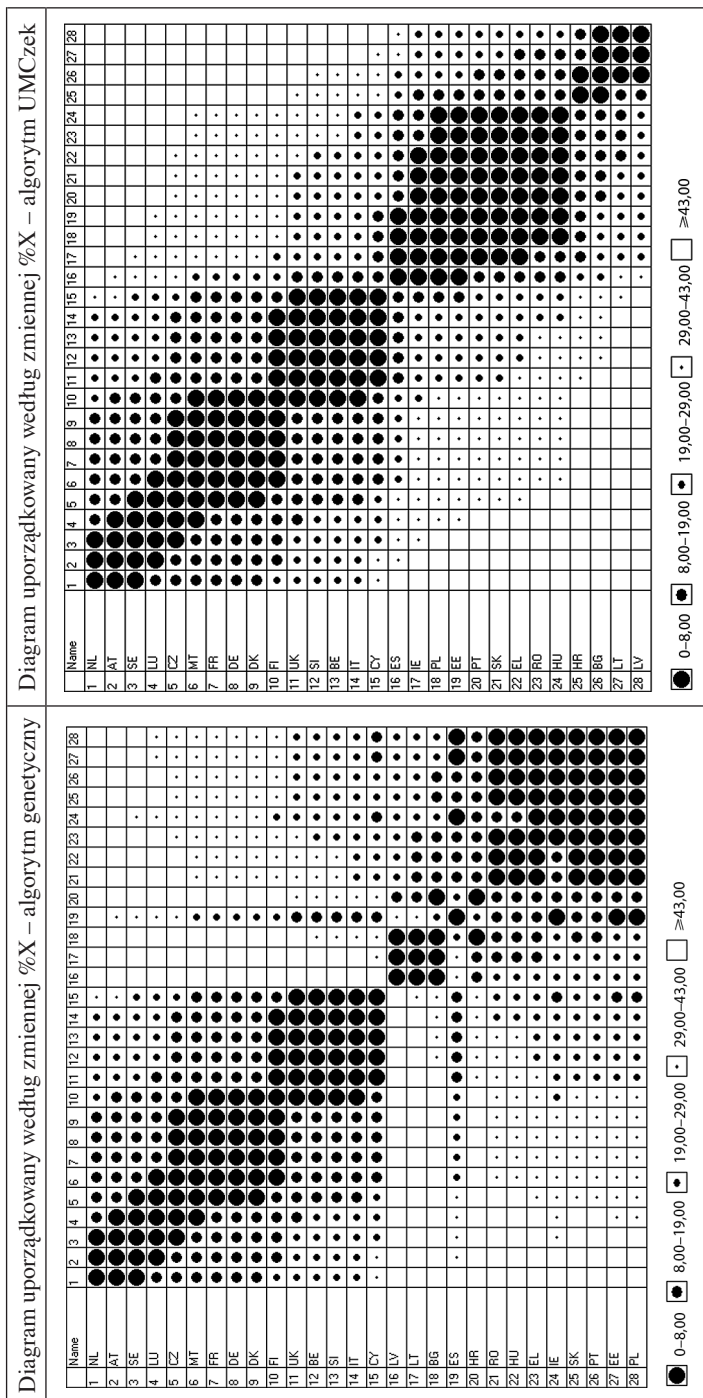








Tabela 12. Grupowanie krajów Europy metodą Czekanowskiego na podstawie realizacji zmiennej %X



Źródło: opracowanie własne.

Jak można zauważyć, w obydwu wariantach wyniki grupowania są bardzo zbliżone, pomimo różnego uporządkowania krajów. Grupę krajów o najkorzystniejszych wartościach zmiennych B-U-D tworzą: Holandia, Austria, Luksemburg, Szwecja i Czechy. Druga grupa krajów o nieco mniej korzystnych wartościach zmiennych diagnostycznych obejmuje 8 krajów: Francję, Niemcy, Wielką Brytanię, Danię, Finlandię, Belgię, Malte i Słowenię. W trzeciej grupie, obejmującej 7 krajów: Włochy, Irlandia, Hiszpania, Portugalia, Polska, Estonia, Cypr – zmienne diagnostyczne kształtują się na poziomie poniżej przeciętnego. W ostatniej grupie znajdują się kraje o najmniej korzystnych wartościach zmiennych. W odróżnieniu od poprzednich grup, kolejność krajów w tej grupie jest identyczna dla obydwu zmiennych: Słowacja, Grecja, Rumunia, Węgry, Chorwacja, Bułgaria, Litwa i Łotwa.

Wyniki grupowania krajów metodą uproszczoną zebrano w tabeli 13, w której także zamieszczono wyniki uzyskane klasyczną metodą Czekanowskiego (ostatnie dwie kolumny). Kolejne segmenty tej tabeli zawierają wyniki otrzymane w uproszczonej metodzie delimitacji dla zmieniającego się poziomu zaokrągleń zmiennej delimitacyjnej począwszy od 30%, gdzie wartości zmiennej zostały zaokrąglone do [0; 0,3; 0,6; 0,9]. Następne poziomy zaokrągleń to 20% z wartościami [0; 0,2; 0,4; 0,6; 0,8] oraz 5% (wartości od 0,10 do 0,90 z odstępem co 0,05). W ostatnim segmencie tej tabeli przytoczono wyniki grupowania na podstawie oryginalnych (niezaokrąglonych) wartości zmiennej diagnostycznej. Wyróżnione grupy krajów o podobnych wartościach zmiennej syntetycznej oznaczono w tabeli różnymi kolorami. Kolorem zielonym oznaczono najmniejsze wartości zmiennej syntetycznej, natomiast kolorem czerwonym – wartości największe. Elementy macierzy zawierające pośrednie wartości zmiennej syntetycznej oznaczono pozostałymi kolorami – od żółtego do niebieskiego. W poszczególnych segmentach tabeli wynikowa liczba grup krajów podobnych (a tym samym i kolorów) jest różna – od 3 grup w pierwszym segmencie (30-procentowy poziom zaokrągleń) do 6 grup w segmencie zawierającym wyniki analizy bez zaokrągleń zmiennej diagnostycznej.

Analizując uzyskane wyniki, należy zwrócić uwagę na wysoką zgodność delimitacji w wariantach opartych na zmiennych standaryzowanych dynamicznie ( $%%X$ ) i statycznie ( $%X$ ). Podobnie jak w przypadku klasycznej metody Czekanowskiego w delimitacji opartej na danych wyjściowych (bez zaokrągleń) kolejność krajów w obydwu wariantach (standaryzacja statyczna i dynamiczna) nie jest identyczna, ale w ramach grup o zbliżonych wartościach zmiennej syntetycznej uzyskuje się pełną zgodność przypisanych do nich krajów.

W przypadku delimitacji opartej na danych zaokrąglonych (do 5%, 20% i 30%) obserwuje się kilka rozbieżnych przyporządkowań. Dotyczy to:

- Finlandii w wariancie z danymi zaokrąglonymi do 5%,
- Słowenii, Wielkiej Brytanii i Hiszpanii w wariancie z danymi zaokrąglonymi do 20%,
- Danii, Malty i Łotwy w wariancie z danymi zaokrąglonymi do 30%.

Tabela 13. Porównanie wyników delimitacji krajów Europy

Lp.	%X	0.3	%%X	0.3	%X	0.2	%%X	0.2	%X	0.05	%%X	0.05	%X	0.05	%%X	0.05	%X	0.0	%%X	0.0	CZ%X	CZ%%X
1	AT	0.0	AT	0.0	NL	0.0	NL	0.0	NL	0.10	NL	0.15	NL	0.14	NL	0.16	NL	0.14	NL	0.16	NL	NL
2	CZ	0.0	CZ	0.0	AT	0.0	AT	0.0	AT	0.15	AT	0.15	AT	0.17	AT	0.20	AT	0.17	AT	0.20	AT	AT
3	LU	0.0	DK	0.0	CZ	0.2	CZ	0.2	SE	0.20	LU	0.20	LU	0.20	LU	0.21	SE	0.20	LU	0.21	SE	LU
4	MT	0.0	LU	0.0	DE	0.2	DE	0.2	LU	0.20	SE	0.20	SE	0.22	SE	0.24	LU	0.22	SE	0.24	LU	SE
5	NL	0.0	NL	0.0	DK	0.2	DK	0.2	CZ	0.25	CZ	0.25	CZ	0.28	CZ	0.28	CZ	0.28	CZ	0.28	CZ	CZ
6	SE	0.0	SE	0.0	FI	0.2	FI	0.2	MT	0.25	DK	0.25	DK	0.30	DK	0.29	MT	0.30	DK	0.29	MT	DK
7	BE	0.3	BE	0.3	FR	0.2	FR	0.2	FR	0.30	DE	0.30	FR	0.30	DE	0.31	FR	0.30	DE	0.31	FR	MT
8	CY	0.3	CY	0.3	LU	0.2	LU	0.2	DE	0.30	MT	0.30	DE	0.31	MT	0.31	DE	0.31	MT	0.31	DE	DE
9	DE	0.3	DE	0.3	MT	0.2	MT	0.2	DK	0.30	FR	0.30	DK	0.32	FR	0.34	DK	0.32	FR	0.34	DK	FI
10	DK	0.3	FI	0.3	SE	0.2	SE	0.2	FI	0.35	FI	0.30	FI	0.36	FI	0.34	FI	0.36	FI	0.34	FI	FR
11	FI	0.3	FR	0.3	UK	0.4	SI	0.2	UK	0.40	SI	0.35	UK	0.40	SI	0.37	UK	0.40	SI	0.37	UK	SI
12	FR	0.3	IT	0.3	SI	0.4	UK	0.2	SI	0.40	UK	0.35	SI	0.41	UK	0.38	BE	0.41	UK	0.38	BE	UK
13	IT	0.3	MT	0.3	BE	0.4	BE	0.4	BE	0.40	BE	0.40	BE	0.41	BE	0.41	SI	0.41	BE	0.41	SI	BE
14	SI	0.3	SI	0.3	IT	0.4	IT	0.4	IT	0.40	IT	0.45	IT	0.42	IT	0.46	IT	0.42	IT	0.46	IT	IT
15	UK	0.3	UK	0.3	CY	0.4	CY	0.4	CY	0.45	CY	0.45	CY	0.46	CY	0.48	CY	0.46	CY	0.48	CY	CY
16	BG	0.6	BG	0.6	ES	0.4	EE	0.6	ES	0.55	EE	0.60	ES	0.57	IE	0.60	ES	0.57	IE	0.60	ES	IE
17	EE	0.6	EE	0.6	EE	0.6	EL	0.6	IE	0.60	ES	0.60	IE	0.63	ES	0.61	IE	0.63	ES	0.61	IE	ES
18	EL	0.6	EL	0.6	EL	0.6	ES	0.6	EE	0.60	IE	0.60	IE	0.64	PT	0.62	PL	0.64	PT	0.62	PL	PT
19	ES	0.6	ES	0.6	HR	0.6	HR	0.6	PL	0.60	PL	0.60	PL	0.64	PL	0.63	EE	0.64	PL	0.63	EE	EE
20	HR	0.6	HR	0.6	HU	0.6	HU	0.6	PT	0.65	PT	0.60	SK	0.65	EE	0.63	PT	0.65	EE	0.63	PT	PL
21	HU	0.6	HU	0.6	IE	0.6	IE	0.6	SK	0.65	SK	0.65	SK	0.65	SK	0.65	SK	0.65	SK	0.65	SK	SK
22	IE	0.6	IE	0.6	PL	0.6	PL	0.6	EL	0.65	EL	0.65	EL	0.70	EL	0.66	EL	0.70	EL	0.66	EL	EL
23	LT	0.6	LT	0.6	PT	0.6	PT	0.6	RO	0.70	RO	0.65	RO	0.70	RO	0.70	RO	0.70	RO	0.70	RO	RO
24	PL	0.6	PL	0.6	RO	0.6	RO	0.6	HU	0.70	HU	0.70	HU	0.70	HU	0.72	HU	0.70	HU	0.72	HU	HU
25	PT	0.6	PT	0.6	SK	0.6	SK	0.6	HR	0.75	HR	0.75	HR	0.78	HR	0.78	HR	0.78	HR	0.78	HR	HR
26	RO	0.6	RO	0.6	BG	0.8	BG	0.8	BG	0.80	BG	0.80	BG	0.84	BG	0.82	BG	0.84	BG	0.82	BG	BG
27	SK	0.6	SK	0.6	LT	0.8	LT	0.8	LT	0.85	LT	0.80	LT	0.87	LT	0.85	LT	0.87	LT	0.85	LT	LT
28	LV	0.9	LV	0.6	LV	0.8	LV	0.8	LV	0.90	LV	0.85	LV	0.90	LV	0.88	LV	0.90	LV	0.88	LV	LV

Źródło: opracowanie własne.

Tabela 14. Zgodność uporządkowań krajów wynikających ze zmiennych syntetycznych %X oraz %%X

Kraj	Miejsce krajów według różnych zmiennych				Moduły różnic miejsc				
	U1	U2	U3	U4	U5	U6	U7	U8	U9
NL	1	1	1	1					
AT	2	2	2	2					
SE	3	4	3	4	1	1	1		
LU	4	3	4	3	1	1	1		
CZ	5	5	5	5					
MT	6	8	6	7	2	2	1		1
FR	7	9	7	10	3	2	3		1
DE	8	7	8	8	1	1			1
DK	9	6	9	6	3	3	3		
FI	10	10	10	9	1		1		1
UK	11	12	11	12	1	1	1		
SI	12	11	13	11	2	1	2	1	
BE	13	13	12	13	1		1	1	
IT	14	14	14	14					
CY	15	15	15	15					
ES	16	17	16	17	1	1	1		
IE	17	16	17	16	1	1	1		
PL	18	19	18	20	2	1	2		1
EE	19	20	19	19	1	1			1
SK	20	21	21	21	1	1		1	
PT	21	18	20	18	3	3	2	1	
EL	22	22	22	22					
HU	23	24	24	24	1	1		1	
RO	24	23	23	23	1	1		1	
HR	25	25	25	25					
BG	26	26	26	26					
LT	27	27	27	27					
LV	28	28	28	28					
Liczba krajów:					U5	U6	U7	U8	U9
zajmujących identyczne miejsca					10	12	15	22	22
różniących się zajmowanymi miejscami					18	16	13	6	6
Liczba krajów różniących się:									
1 miejscem					12	12	8	6	6
2 miejscami					3	2	3		
3 miejscami					3	2	2		
Suma modułów różnic miejsc					27	22	20	6	6

Oznaczenia parametrów:

U1 – %X; U2 – %%X; U3 – CZ%X; U4 – CZ%%X; U5 – max(U1:U4) – min(U1:U4);

U6 – moduł (U1 – U2); U7 – moduł (U3 – U4); U8 – moduł (U1 – U3); U9 – moduł (U2 – U4).

Źródło: opracowanie własne.



Identyfikatory nazw krajów o rozbieżnych przyporządkowaniach oznaczono w tabeli 13 pogrubioną czcionką.

Nieco inaczej jest, jeżeli chodzi o zgodność grup krajów w przekroju wariantów analizy różniących się stopniem zaokrąglenia danych. W tym przypadku pomimo generalnej zgodności uporządkowań krajów (por. tabela 14) uzyskuje się różne konfiguracje i różne liczby grup krajów. W wariancie zaokrąglenia 30% otrzymano 3 ugrupowania, w wariancie z danymi zaokrąglonymi do 20% – 5 grup, natomiast w pozostałych dwóch wariantach (zaokrąglenia 5% oraz bez zaokrąglenia) pojawia się podział na 6 grup. Z klasycznej metody Czekanowskiego wynika jeszcze inna konfiguracja z 4 grupami.

Z informacji podanych w tabeli 14 wynika, że największa zgodność uporządkowań ma miejsce w przypadku kolejności wynikającej z klasycznej metody Czekanowskiego oraz zmiennych %X i %%X (dwie ostatnie kolumny tej tabeli). Jak można zauważyć, dla 22 krajów ich kolejność w tych metodach delimitacji się pokrywa. Nieco większe rozbieżności wystąpiły w uporządkowaniach krajów według zmiennych standaryzowanych (statycznie i dynamicznie). Liczba krajów o zgodnych pozycjach spada tu z 22 do 12–15. Zmiany w uporządkowaniach miały miejsce m.in. dla Francji, Danii, Portugalii, Malty, Słowenii i Polski.

Ogólnie biorąc, różnice te są niewielkie, a ponadto w trakcie grupowania większość z nich zanika, gdyż do danej grupy są zaliczane kraje o zbliżonych pozycjach wynikających ze zmiennych delimitacyjnych.

## 5. Wnioski

Wyniki zastosowania uproszczonej metody delimitacji wskazują na wysokie podobieństwo ugrupowań krajów w przekroju cech standaryzowanych dynamicznie i statycznie. W dalszych badaniach warto podjąć próbę ustalenia, czy jest to związane z małymi zmianami cech w analizowanych latach, czy też decyduje o tym algorytm standaryzacji.

Uzyskiwanie różnych wyników grupowania w zależności od stopnia zaokrąglenia danych wyjściowych może świadczyć o przydatności tego narzędzia w analizach. Jeżeli rezultaty analizy są identyczne, to stosowanie tej operacji jest bezzasadne.

Operacja zaokrąglania danych w analizach danych właściwie nie jest stosowana. Jak się wydaje, w niektórych sytuacjach może się ona okazać przydatna. Pozwala przyglądać się dużym zbiorom liczb z pewnym (stopniowanym) dystansem, łatwiej jest wówczas dostrzec istniejące w nich regularności. Inaczej mówiąc, zaokrąglanie

liczb to narzędzie numerycznej indukcji umożliwiające analizy danych od szczegółu do ogółu.

Do zalet uproszczonych metod delimitacji należy też zaliczyć ich prostotę, intuicyjność, możliwość śledzenia efektów analizy w każdym jej etapie, a także brak ograniczeń związanych z narzędziami analitycznymi.

Prezentowana metoda delimitacji opiera się na pojedynczych zmiennych, co może być uznane za jej wadę. Tym niemniej jeżeli analizowana cecha jest zmienną agregatową, to można przyjąć, że pozwala ona także analizować struktury wielowymiarowe.

Istotnym problemem jest kwestia wykorzystania w analizach danych wyraźnych w różnych skalach pomiarowych, a zwłaszcza zasadności wykonywania na zmiennych w skalach słabszych: porządkowej (rangi) lub nominalnej (kategorii) operacji i przekształceń stosowanych w przypadku zmiennych w skalach silnych (ilorazowa lub przedziałowa). Warto podjąć badania mające na celu ustalenie, jak dalece zmieniają się wyniki analiz (np. parametry współzależności) w przypadku zastąpienia zmiennych w skalach silnych zmiennymi w skalach słabszych.

## Literatura

- Czekanowski J. [1913], *Zarys metod statystycznych w zastosowaniu do antropologii*, Prace Towarzystwa Naukowego Warszawskiego, nr 5.
- Kozłowski S. [1972], *O sposobie weryfikacji podziałów przestrzennych w odniesieniu do taksonomicznej metody różnic J. Czekanowskiego*, „Wiadomości Statystyczne”, nr 1.
- Piasecki Z. [1971], *Nowa metod taksonomiczna*, „Listy Biometryczne”, nr 30–31.
- Sołtysiak A. [1977a], *UMCzek 1.00. Ewolucyjny algorytm porządkowania diagramu Czekanowskiego*, „Biuletyn Antropologiczny”, t. 1.
- Sołtysiak A. [1977b], *MaCzek 1.00. Język programowania diagramu Czekanowskiego*, „Biuletyn Antropologiczny”, t. 1.

## The Simplified Method of Vector Delimitation

(Abstract)

The paper presents a graphical procedure for sorting and grouping objects described by a single variable taken in the form of cross-sectional, time series or panel data. This method draws on the classical taxonomic method developed over 100 years ago by Jan Czekanowski. The advantages of simplified vector delimitation are that it is not labour intensive but is intuitive and can be handled with simple spreadsheet functions. The method was used for the taxonomic analysis of a set of 28 EU countries described by three characteristics: the unemployment rate, poverty risk and life expectancy for the two years 2010 and 2015.

**Keywords:** taxonomy, clustering, grouping, delimitation, cross-time data.