UNIWERSYTET SZCZECIŃSKI
WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA

Studia i Prace WNEIZ US
nr 47/3 2017

PROBLEMY TEORETYCZNE I METODYCZNE

DOI: 10.18276/SIP.2017.47/3-03

Bogusław Kaczmarczyk*
Łukasz Popławski*
Uniwersytet Ekonomiczny w Krakowie, Wydział Finansów

TAKSONOMICZNA ANALIZA INFORMACYJNOŚCI KOSTEK DANYCH – OBJĘTOŚĆ INFORMACYJNA NA WYBRANYM PRZYKŁADZIE EKOROZWOJU

STRESZCZENIE

Termin objętość informacyjna zawarty w tytule stanowi hasło wywoławcze dla dość obszernej nowej problematyki związanej z taksonomią i klasyfikacją obiektów bądź kostek danych w ujęciu regionalnym. W analizach danych koncentracja badawcza skupiona jest głównie na zmiennych, a rozwój metod, technik i narzędzi analizy danych w tym względzie jest ogromny. Celem artykułu jest prezentacja podstaw dla grupowania obiektów pod względem analizy poziomej kostki danych z wykorzystaniem pojęcia objętości informacyjnej w związku z ilościowym zagadnieniem bliskości zarówno obiektów w kostce, jak i możliwości analiz metrycznych kostek względem siebie. W pracy jako studium przypadku grupowania obiektów wykorzystano dane dla wybranych gmin obszaru województwa świętokrzyskiego w związku z ich ekorozwojem.

Słowa kluczowe: objętość informacji, kostka danych, ekorozwój

* Adres e-mail: b.kaczmarczyk@pro.onet.pl.
** Adres e-mail: rmpoplaw@gmail.com.

Wprowadzenie

Objętość informacyjna dotyczy każdorazowo zbiorów mierzalnych w przestrzeni (*nk*) + 1 wymiarowej dla szeregu obiektów O_i opisanych zmiennymi x_i. Wyjściowy zbiór danych stanowi macierz X_i jako kostka danych:

$$X_i = \{O_i = \{x_1, x_2, ..., x_n\}\} \subset X_i \qquad \textbf{(1)}$$

$$X_i \Rightarrow X_{n,k} = \begin{matrix} O_1 \\ O_2 \\ O_3 \\ \dots \\ O_n \end{matrix} \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,k} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,k} \\ x_{3,1} & x_{3,2} & x_{3,3} & \dots & x_{3,k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n,1} & x_{n,2} & x_{n,3} & \dots & x_{n,k} \end{bmatrix} \qquad \textbf{(1.1)}$$

W tym miejscu warto wskazać na geometryczną postać i zarazem własność wzajemnie jednoznacznego przyporządkowania różnym obiektom O_i różnych odległości d_{O_I} i wzajemnie odwrotnie przyporządkowanie różnym odległościom różnych obiektów w mierzalnych kostkach danych $X_{n,k}$, kostkach mogących przyjmować jeden z trzech wymiarów:

a) jeżeli $n > k$ $V_{\text{inf.}}$ dotyczy prostokątnej i pionowej kostki – **układ nadokreślony** $X_{n>k}$, np. $X_{5,3}$:

$$X_{5,3} = \begin{matrix} O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \end{matrix} \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,2} & x_{3,3} \\ x_{4,1} & x_{4,2} & x_{4,3} \\ x_{5,1} & x_{5,2} & x_{5,3} \end{bmatrix} \xleftrightarrow{Odpowiada} \begin{bmatrix} d_{O_1} \\ d_{O_2} \\ d_{O_3} \\ d_{O_4} \\ d_{O_5} \end{bmatrix}$$

b) jeżeli $n = k$ $V_{\text{inf.}}$ obejmuje kwadratową kostkę danych – **układ tożsamy** $X_{n=k}$, np. $X_{2,2}$

$$X_{2,2} = \begin{matrix} O_1 \\ O_2 \end{matrix} \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \end{bmatrix} \longleftrightarrow \text{Odpowiada} \longleftrightarrow \begin{bmatrix} d_{O_1} \\ d_{O_2} \end{bmatrix}$$

c) jeżeli $n < k$ $V_{\text{inf.}}$ związana jest z prostokątną i poziomą kostką danych – **układ niedookreślony**[1] $X_{\text{n<k}}$, np. $X_{3,5}$

$$X_{3,5} = \begin{matrix} O_1 \\ O_2 \\ O_3 \end{matrix} \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & x_{1,5} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & x_{2,5} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} & x_{3,5} \end{bmatrix} \longleftrightarrow \text{Odpowiada} \longleftrightarrow \begin{bmatrix} d_{O_1} \\ d_{O_2} \\ d_{O_3} \end{bmatrix}$$

1. Krótki opis teoretyczny objętości informacyjnej w kostkach danych

Objętość informacji $V_{\text{inf.}}$ wyznacza się za pomocą przyjętej odległości w przestrzeni wielowymiarowej. Według [Coombsa, Dawesa, Tversky'ego – (1977, s. 97)], [Jajugi (1993, s. 56)] i wielu innych szczególną klasę odległości stanowi przypadek metryki Minkowskiego dla $r = 2$, czyli odległości Euklidesa ozn. w tekście jako[2] $\|O_i\|_2$. W większości kostek danych $X_{i.}$ dla każdego z obiektów O_i (wektorów informacji) odległość Euklidesa jest przypisana w sposób prawie na pewno wzajemnie jednoznaczny (izomorfizm), bądź co najmniej jednoznaczny (zależność homomorficzna). Zatem formalnie dla opisu objętości informacyjnej $V_{\text{inf.}}$ na kostce $X_{i.}$ wskazano dwie z pięciu definicji opisowych[3] związanych z $V_{\text{inf.}} X_{i.}$:

– *definicja pierwsza*: „w rozpiętym na obiektach O_n wielowymiarowym układzie współrzędnych *WUW* dla mierzalnej, kompletnej i wielowymiarowej

[1] W dziedzinie nauk ekonomicznych, w dyscyplinie finanse w skali mikro danych, przypadek **prostokątny i poziomy** w $X_{i_{n,k}}$ występuje bardzo często w praktyce. Na gruncie problematyki regresyjnej jako problem Gaussa-Markowa, mierzalny **układ prostokątny i poziomy**, dla którego liczba wierszy (obiektów) **jest mniejsza** od liczby kolumn (zmiennych), posiada w klasie rozwiązań liniowych z wykorzystaniem macierzy **MP-odwrotnych** jednoznaczne rozwiązanie. Szerzej zob. [Kaczmarczyk, 2015, s. 115–162, mat. niepublikowany].

[2] Odległość Euklidesa jako pierwiastek drugiego stopnia z sumy różnic kwadratów dla poszczególnych współrzędnych kostki danych, synonim *norma Euklidesa* ozn. $\|O_i\|_2$.

[3] Definicje autora (B. Kaczmarczyk). Pozostałe własności objętości informacji, również definicje trzecia, czwarta i piąta, zostaną zaprezentowane w części drugiej artykułu.

kostki danych X_i objętością informacyjną $V_{\text{inf.}} X_i$ jest wielokrotny (n-krotny) iloczyn długości $||O_n||_2$ wszystkich jej obiektów O_n". Zapis formalny dla *definicji pierwszej* $V_{\text{inf.}} X_i$

$$V_{\text{inf.}} X_i = \prod_{1}^{n} \|O_n\|_2 \qquad \textbf{(2)}$$

$$d_{O_i} = \|O_i\|_2 \qquad \textbf{(2.1)}$$

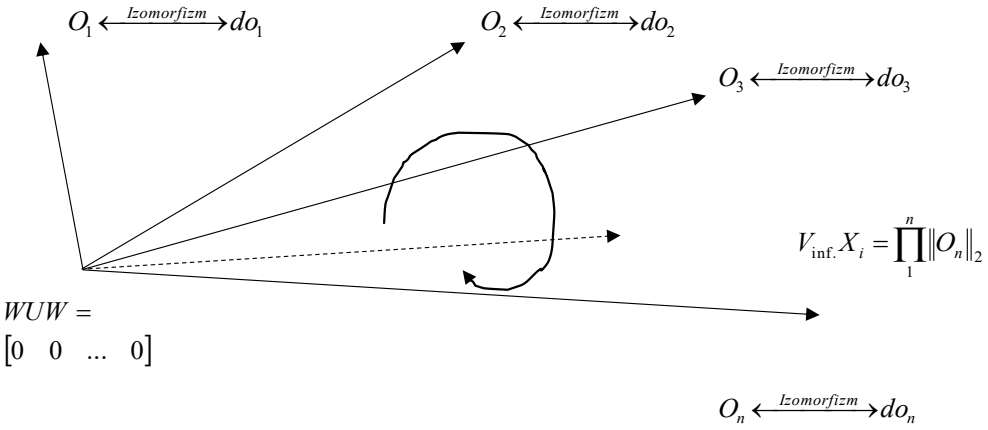
gdzie $d_{O_i} = ||O_n||_2$ oznacza przypisaną dla wszystkich obiektów O_n długość Euklidesa liczoną od początku zapiętego wielowymiarowego układu współrzędnych[4] w ramach $R^{(nk)+1}$. Wskazana w *definicji pierwszej* i formule **(2)** objętość informacja $V_{\text{inf.}} X_i$ w ujęciu izomorficznym ma swoją interpretację geometryczną[5] dla wzajemnie jednoznacznego przyporządkowania w przestrzeni $R^{(nk)+1}$ wymiarowej[6] jako „miotełka obiektów" iloczynu długości w maksymalnym wymiarze kostki danych $X_{i.}$ – rysunek 1.

– *definicja druga (prawie na pewno izomorfizm)*: „przydział obiektom O_i metryk, tj. długości d_{o_i} w ramach $V_{\text{inf.}}$ w $X_{i.}$ jest na ogół izomorficzny ze względu na stosunek długości odcinków. Relacja R odwzorowuje zatem **izomorficznie** (wzajemnie jednoznacznie) stosunek długości między obiektami w ramach kostki danych na stosunku większości pomiędzy liczbami jako długościami tych obiektów". Dla *definicji drugiej* **zasada odwrotna (tj. zasada homomorfizmu) może być spełniona,** bowiem jest możliwe dla $V_{\text{inf.}}$ przyporządkowanie jednoznaczne w ramach tej samej kostki danych **jednej długości kilku różnym obiektom** w $X_{i.}$.

[4] W analizach wielowymiarowych przyjęcie początku wielowymiarowego układu współrzędnych dla wszystkich n obiektów opisanych liczbą k zmiennych zwiększa wymiar analizy o jeden.

[5] Interpretacja w sensie geometrii obliczeniowej.

[6] Wymiar $R^{(nk)+1}$ uwzględnia zapięcie wielowymiarowego układu współrzędnych w ramach „*WAP*".



Rysunek 1. Objętość informacja $V_{inf.}\, X_i$ jako izomorficzna „miotełka obiektów" w przestrzeni $R^{(nk)+1}$ wymiarowej z wykorzystaniem odległości Euklidesa przypisanej wzajemnie jednoznacznie obiektom O_i

Źródło: opracowanie własne.

2. Taksonomiczna analiza informacyjności kostek danych – objętość informacyjna w ujęciu niemnogościowym

W przypadku objętości informacji na kostce danych $V_{\text{inf.}} X_i$ kryterium podziału odnosi się początkowo do długości $d_{O_i} = ||O_n||_2$ danego obiektu jako wektora informacji od początku zapiętego układu współrzędnych, a końcowo do części wspólnej tych długości, tj. do iloczynu (analiza pozioma, tzw. analiza obiektowa kostki danych) dla czynności porządkowania obiektów opisanych zmiennymi. Przykładowy algorytm wyznaczania objętości informacyjnej $V_{inf.}$ jako **liniowej miary podobieństwa** dla kostki danych X_i w funkcji jej objętości, z wykorzystaniem wzajemnych odległości Euklidesa[7] w sensie izomorficznym pomiędzy obiektami, przedstawia poniższa tabela 1, wraz z kierunkami dalszych badań – tabela 2.

[7] Odległość Euklidesa, oznaczona jako $d_{O_i}\ ||O_i||_2$ poza odległościami: Czebyszewa, Minkowskiego, miejską, Mahalanobis, Czekanowskiego i inne, stanowi jedną z wielu możliwych odległości ogólnie stosowanych w ramach *WAP*.

Tabela 1. Metodyka taksonomicznego wyznaczania objętości informacyjnej $V_{inf.}$ jako funkcji porządkującej liniowo badane obiekty w ramach kostki danych X_i – ujęcie izomorficzne

Lp.	Czynność	Etap
1.1.	Zebranie kompletnych mianowanych i mierzalnych danych i postawienie problemu klastrowania	I Zebranie i przygotowanie danych do analizy wielu zmiennych
1.2.	Analiza jakościowa danych, analizy ilościowe, w tym obliczenie macierzy korelacji dla zmiennych i wnioskowanie w zakresie wyjściowego nieunormowanego zbioru zmiennych	
1.3.	Wyjściowa kostka danych X_i	
1.4.	Dokonanie transformacji cech zmiennych związanej z wyeliminowaniem jednostek i dominacji zmiennych poprzez zniesienie efektu skali. Proces ten dotyczy: normalizacji, standaryzacji, unitaryzacji, stosowania metod: rangowych, metod ilorazowych, metody T. Grabińskiego bądź zastosowanie innych metod transformacji kostki danych X_i z uwzględnieniem natury analizowanego zjawiska, własności i charakteru zmiennych	
2.1.	Zapięcie wielowymiarowego układu współrzędnych WUW = [0 0 … 0] na obiektach O_i przetransformowanej kostki danych X_i	II Analiza danych i procedura grupowania objętościowego kostki danych
2.2.	Obliczenie sumy kwadratów dla wektorów informacji w kostce danych	
2.3.	Obliczenie długości wektorów (obiektów) informacji w kostce danych d_{O_i} $\|\|O_i\|\|_2$	
2.4.	Obliczenie procentowego udziału informacji dla danego wektora w kostce danych	
2.5.	Obliczenie skumulowanego procentu informacji danego wektora w kostce danych	
2.6.	Obliczenie objętości informacyjnej analizowanej kostki danych $V_{\text{inf.}} X_i$ na podstawie d_{O_i} w ujęciu izomorficznym $\begin{bmatrix} d_{O_1} = \|O_1\|_2 \\ d_{O_2} = \|O_2\|_2 \\ d_{O_3} = \|O_3\|_2 \\ \dots \\ d_{O_n} = \|O_n\|_2 \end{bmatrix} \Rightarrow V_{\text{inf.}} X_i = \prod_1^n \|O_i\|_2$ [d] *definicja pierwsza* i formuła (2)	
2.7.	Przyjęcie kryterium podziału i sporządzenie diagramu nieuporządkowanego objętościowo badanych obiektów w ramach kostki danych X_i w ramach $n!$ możliwych podzbiorów kostki danych	
2.8.*	Orientacja cech dla zmiennych w kostce danych. Obliczenie kosinusów kierunkowych dla obiektów i zmiennych x_i w wielowymiarowym układzie współrzędnych w ramach kostki danych X_i $\cos\alpha_{O_1,x_1} = \frac{x_1}{\|O_1\|_2}; \cos\alpha_{O_1,x_2} = \frac{x_2}{\|O_2\|_2}, \dots, ; \cos^2\alpha_{O_1,x_1} + \dots + \cos^2\alpha_{O_n,x_n} = 1$	
2.9.*	Sporządzenie diagramu uporządkowanego objętościowo, delimitacja kostki danych	

Lp.	Czynność	Etap
2.10.*	Uzupełnienie analizy w celu orientacji obiektowej poprzez wyznaczenie kątów pomiędzy obiektami O_i i O_j w ramach kostki danych X_i $\cos\psi_{O_i,O_j} = \frac{O_i \circ O_j}{\|O_i\|_2 \cdot \|O_j\|_2}$ symbol O oznacza mnożenie skalarne wektorów informacji	II Analiza danych i procedura grupowania objętościowego kostki danych
2.11.*	Możliwe obliczenie i wyprowadzenie odległości kątowej pomiędzy obiektami dla przyjętego układu współrzędnych	
2.12.*	Ujęcie dynamiczne analizy kostek i obliczenie charakterystyk w tym zakresie; *oznacza dodatkową możliwość, tj. ujęcie dynamiczne analizy danych w dwóch stanach badawczych jak dla metody różnicowej analizy danych. Analiza zbiorów o skończonej liczbie elementów, zbiorów o równej mocy	
2.13.	Grupowanie objętościowe, profil obiektów z możliwą wizualizacją struktury danych z wykorzystaniem metod i narzędzi geometrii obliczeniowej, analiza wyników	
2.14.	Ujęcie mnogościowe (teoria zbiorów) objętości informacji $V_{\text{inf.}}$	
3.1.	Wnioski końcowe	III Wnioskowanie

Źródło: opracowanie własne.

Kierunki dalszych badań dla $V_{\text{inf.}}$ – tabela 2.

Tabela 2. Kierunki dalszych badań i prac nad $V_{inf.}$ w ramach kostki danych X_i

Lp.	Kierunek dalszych badań nad $V_{\text{inf.}}$	Etap
1.	Nowe hipotezy badawcze, przykładowo: H_1: czy $V_{inf.}$ może stanowić kategorie dla metod wzorcowych porządkowania liniowego? H_2: czy istnieje oraz jaka jest odporność $V_{inf.}$ w ramach *WAP* na przyjęte i stosowane w nauce skale pomiarowe? w tym hipoteza pomocnicza: $H_{2.1}$: Jaki jest wpływ transformacji cech dla porządku klastrowania objętościowego kostek danych? H_3: czy dla $V_{inf.}$ istnieje pomiar jakościowej kostki danych?	Nowe hipotezy

Źródło: opracowanie własne.

3. Studium przypadku zastosowania objętości informacji $V_{inf.}$ do klasyfikacji obiektów na podstawie danych dla wybranych gmin obszaru województwa świętokrzyskiego

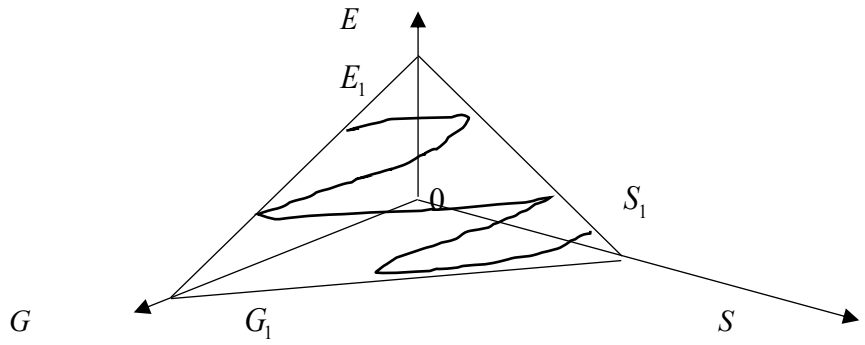
Rozważmy jako studium przypadku[8] przykład zaczerpnięty z pracy Popławskiego, (2009) w odniesieniu do zmiennych ekologicznych oznaczonych jako typ *E* w ramach pojęcia ekorozwój – rysunek 2.

Rysunek – 2 przedstawia ekorozwój w znaczeniu podwójnym:

– po pierwsze według Zaufala – (1983), Górki – (2007) jako wzrost gospodarczy zgodny z wymogami ochrony środowiska życia człowieka, w tym zwłaszcza ochrony przyrody,

– po drugie jako „EKRW" – jako płaszczyznę wspólną pojęć[9]: ekologia „E", gospodarka „G" i społeczeństwo „S" z uwzględnieniem warunku **(3)**:

$$EKRW = E \cup S \cup G \xrightarrow{Odpowiada} F(E,S,G) = \frac{E_1}{E} + \frac{S_1}{S} + \frac{G_1}{G} = \overline{1} \qquad \textbf{(3)}$$



Rysunek 2. Ekorozwój w znaczeniu łącznym

Źródło: opracowanie własne.

W przykładzie celem analizy jest dokonanie przestrzennego grupowania wybranych gmin województwa świętokrzyskiego, wchodzących w skład Nadnidziańskiego

[8] Wykorzystano dane zawarte w pracy Popławskiego (2009, s. 205 i nast.).

[9] Parametry E_1, S_1, G_1 oznaczają konkretne dane ekorozwoju.

Parku Krajobrazowego, za pomocą $V_{\text{inf.}}$. Uwzględniając punkty 1.1. do 1.3. powyższej tabeli 1, po analizie[10] ilościowo-jakościowej otrzymano dane zestawione[11] w tabeli 3.

Tabela 3. Dane do analizy grupowania

Nazwa gminy	X27	X28	X44	X57	X69	Obszar gminy (w km^2)	Ludność według stanu na 2006r.	Dochód ogółem na osobę (w tys. zł)
Opis zmiennych typu *E*	Udział obszarów prawnie chronionych w powierzchni ogólnej	Udział lesistości w powierzchni ogólnej	Wskaźnik lokalizacji przestrzennej	Długość sieci kanalizacyjnej w km na 1 km^2	Udział gruntów ornych w gospodarstwach indywidualnych (w %)	Dane geograficzno-demograficzne		Dane GUS na podstawie sprawozdań branżowych samorządów typu Rb-27S
Imielno	0,5116	0,1060	0,1675	0,0000	72,9609	100,6	4 626	1,9564
Kije	0,9509	0,1784	0,2130	0,0030	61,1231	99,26	4 692	3,9842
Michałów	1,0076	0,1984	0,2133	0,0000	80,3717	112,21	4 839	1,9512
Nowy Korczyn	0,9685	0,0702	0,2487	0,0000	65,8673	117,3	6 381	2,5261
Opatowiec	0,9597	0,1250	0,2056	0,0145	76,3250	68,41	3 599	1,6090
Wiślica	0,9752	0,0324	0,3180	0,0000	64,7167	100,6	5 690	2,8707
Złota	0,9100	0,1243	0,2833	0,3902	72,8060	81,7	4 877	2,4822
Parametry opisowe								
Suma	6,2835	0,8347	1,6493	0,4077	494,1708	680,08	34 704	17,3798
Mediana	0,9597	0,1243	0,2133	0	72,806	100,60	4 839	2,4822
Wartość średnia	0,90	0,12	0,24	0,06	70,60	97,15	4 957,71	2,48
Odchylenie standardowe	0,17	0,06	0,05	0,15	6,90	16,96	876,48	0,79
Zmienność cechy	19%	48%	22%	251%	10%	17%	18%	32%
x Min	0,5116	0,0324	0,1675	0	61,1231	68,41	3 599	1,609
x Max	1,0076	0,1984	0,3180	0,3902	80,3717	117,3	6 381	3,9842
R = x Max – - x Min	0,4960	0,1660	0,1505	0,3902	19,2486	48,89	2782	2,3753
Środek ciężkości zbioru wielocechowego	0,90	0,12	0,24	0,06	70,60	97,15	4 957,71	2,48

Źródło danych: Popławski (2009, s. 337–341).

[10] Analiza korelacyjna i analiza merytoryczna ze względu znaczenie zmiennych w ekorozwoju gmin.

[11] Wynik analizy jakościowej i analizy korelacji dla pełnej macierzy danych $X_{i_{n,k}}$.

W ekonometrii w ramach *WAP* obliczenia i wnioskowanie bez transformacji cech na ogół tracą wartość poznawczą[12] analizowanego zagadnienia. Dlatego w taksonomii, w zależności od typu i własności skali pomiaru (Stevens, 1946, s. 677–680; Ackoff, 1969), s. 243–244, tablica 6.3 i 6.4 klasyfikacja skal pomiaru: nominalna, porządkowa, interwałowa, ilorazowa) dla zmiennych opisanych cechami, istnieje wiele sposobów transformacji, wśród których można wskazać: normowanie, standaryzowanie, unitaryzację, unitaryzację zerowaną i przekształcenia ilorazowe. Zatem przygotowując dane do dalszych obliczeń, w celu eliminacji efektu skali, wyeliminowania różnych jednostek oraz doprowadzania danych do porównywalności, dokonano zgodnie z pkt 1.4. tabeli 1 transformacji cech jednolicie dla wszystkich zmiennych w ramach $\bar{X}$ według formuły **(4)**:

$$z_{i,j} = \frac{x_{i,j} - \bar{X}}{S(X_j)}; \forall S(X_j) > 0 \qquad \textbf{(4)}$$

gdzie:
$Z_{i,j}$ – zmienna standaryzowana,
$X_{i,j}$ – zmienna nieunormowana,
$\bar{X}$ – wartość średnia zmiennej nieunormowanej,
$S(X_j)$ – odchylenie standardowe zmiennych nieunormowanych.

Wyniki standaryzacji cech w ramach $X_{i_{n,k}}$ dla formuły **(4)** zestawiono w tabeli 4.

Kolejnym krokiem było zapięcie wielowymiarowego układu współrzędnych $WUW = [0\ 0\ \ldots\ 0]$ dla unormowanych danych na kostce $X_{i_{n,k}}$ oraz realizacja punktów 2.1. do 2.6. tabeli 1. Wyniki obliczeń zestawiono w tabeli 5.

[12] Przykładowo z powodu efektu rzędu wielkości cech dla zmiennych wyrażonych w różnych jednostkach pomiaru w ramach $X_{i_{n,k}}$.

Tabela 4. Dane unormowane

Wyszczególnienie	X27	X28	X44	X57	X69	Obszar gminy (w km^2)	Ludność według stanu na 2006 r.	Dochód ogółem wykonany na osobę (w tys. zł)
Imielno	–2,2351	–0,2289	–1,3290	–0,3976	0,3428	0,2032	–0,3785	–0,6676
Kije	0,3084	1,0263	–0,4412	–0,3771	–1,3732	0,1242	–0,3032	1,9042
Michałów	0,6364	1,3729	–0,4346	–0,3976	1,4171	0,8877	–0,1354	–0,6743
Nowy Korczyn	0,4105	–0,8510	0,2542	–0,3976	–0,6855	1,1879	1,6239	0,0549
Opatowiec	0,3593	0,0993	–0,5849	–0,2987	0,8305	–1,6949	–1,5502	–1,1083
Wiślica	0,4488	–1,5059	1,6062	–0,3976	–0,8522	0,2032	0,8355	0,4920
Złota	0,0716	0,0873	0,9294	2,2663	0,3204	–0,9112	–0,0921	–0,0008
Suma	0							
Wartość średnia	0							
Odch. stand.	1							
Środek ciężkości zbioru wielocechowego	0							

Źródło: opracowanie własne.

Tabela 5. Objętość informacyjna kostki danych $V_{inf.}X_i$

Nazwa gminy	Sumy kwadratów dla wektorów informacji w kostce danych	Długości wektorów informacji w kostce danych	Procent informacji danego wektora w kostce danych X_i	Procent informacji skumulowany danego wektora w kostce danych	Objętość informacyjna kostki danych $V_{\text{inf.}}X_i$
Zapięcie $WUW = [0\ \ 0 \ldots 0]$					
Imielno	7,7201	2,78	15,2	15,2	$\begin{bmatrix} d_{O_1} = \|O_1\|_2 \\ d_{O_2} = \|O_2\|_2 \\ d_{O_3} = \|O_3\|_2 \\ \ldots \\ d_{O_n} = \|O_n\|_2 \end{bmatrix} \Rightarrow V_{\text{inf.}}X_i = \prod_1^n \|O_i\|_2 = 825{,}6$ **= 825,6**
Kije	7,1044	2,67	14,6	29,7	
Michałów	5,9063	2,43	13,3	43,0	
Nowy Korczyn	5,6362	2,37	13,0	56,0	
Opatowiec	7,7641	2,79	15,2	71,2	
Wiślica	6,9148	2,63	14,4	85,6	
Złota	6,9541	2,64	14,4	**100**	
Suma	48,0	18,30	**100**		

Źródło: opracowanie własne.

Tabela 6. Tabela nieuporządkowana kostki danych dla istniejącej $V_{inf.}X_i$

Nazwa gminy	Cosinusy kierunkowe dla X27	Cosinusy kierunkowe dla X28	Cosinusy kierunkowe dla X44	Cosinusy kierunkowe dla X57	Cosinusy kierunkowe dla X69	Cosinusy kierunkowe dla zmiennej obszar gminy	Cosinusy kierunkowe dla zmiennej ludność	Cosinusy kierunkowe dla zmiennej dochód ogółem wykonany na osobę	Nieuporządkowanie objętościowe kostki danych, objętości cząstkowe kostki danych	Lp.
Imielno	–0,8044	–0,0824	–0,4783	–0,1431	0,1234	0,0731	–0,1362	–0,2403	125,3	1
Kije	0,1157	0,3850	–0,1655	–0,1415	–0,5152	0,0466	–0,1137	0,7144	120,2	2
Michałów	0,2619	0,5649	–0,1788	–0,1636	0,5831	0,3653	–0,0557	–0,2775	109,6	3
Nowy Korczyn	0,1729	–0,3585	0,1071	–0,1675	–0,2887	0,5003	0,6840	0,0231	107,1	4
Opatowiec	0,1290	0,0356	–0,2099	–0,1072	0,2981	–0,6083	–0,5563	–0,3978	125,7	5
Wiślica	0,1707	–0,5727	0,6108	–0,1512	–0,3241	0,0773	0,3177	0,1871	118,6	6
Złota	0,0271	0,0331	0,3524	0,8594	0,1215	–0,3456	–0,0349	–0,0003	119,0	7
Suma									825,6	

Źródło: opracowanie własne.

Przy kontynuacji obliczeń z tabeli 1 kolejny etap stanowi zestawienie odpowiednio nieuporządkowanej i uporządkowanej tabeli danych względem $V_{\text{inf.}}X_i$. Wyniki zaprezentowano w tabeli 6, a delimitację kostki danych w tabeli 7.

Tabela 7. Tabela delimitacji kostki danych dla istniejącej $V_{\text{inf.}}X_i$

Uporządkowanie objętościowe kostki danych. Objętości cząstkowe kostki danych. Profil obiektów względem zastosowanych zmiennych	Lp. sortowana	Nazwa gminy, obiekt wielowymiarowej kostki danych	Procent objętości informacji danego wektora w kostce danych X_i	Procent objętości informacji skumulowany danego wektora w kostce danych
Zapięcie WUW = [0 0 … 0]				
107,1	4	Nowy Korczyn	13,0	13,0
109,6	3	Michałów	13,3	26,3
118,6	6	Wiślica	14,4	40,6
119,0	7	Złota	14,4	55,0
120,2	2	Kije	14,6	69,6
125,3	1	Imielno	15,2	84,8
125,7	5	Opatowiec	15,2	**100**
825,6		Suma	**100**	

Źródło: opracowanie własne z wykorzystaniem funkcji Excela: formatowanie warunkowe.

Podsumowanie

Za pomocą trzyetapowego algorytmu związanego z realizacją klastrowania obiektów izomorficznych, w ramach kostki danych na podstawie objętości informacyjnej $V_{\text{inf.}} X_i$, uzyskano dla wyjściowych danych wyodrębnienie podzbiorów, tzw. taksonów, którym wzajemnie jednoznacznie odpowiadają zarówno odległości, jak i objętości cząstkowe o najmniejszym zróżnicowaniu w ramach łącznej objętości informacji rozważanej kostki.

W ramach gmin wchodzących w skład Nadnidziańskiego Parku Krajobrazowego oraz przyjętego kryterium podziału opartego na pojęciu objętości informacyjnej $V_{\text{inf.}}$ jako kategorii porządkowania liniowego uzyskano rozbicie analizowanej kostki danych X_i na trzy podzbiory: pierwszy {4,3} = {Nowy Korczyn, Michałów}, drugi {6, 7, 2} = {Wiślica, Złota, Kije}, trzeci {1,5} = {Imielno, Opatowiec}.

Kierunkiem dalszych badań, poza formalną stroną zagadnienia $V_{\text{inf.}} X_i$ – tabela 2, w ramach pojęcia ekorozwój, jest kwantyfikacja oparta dla $V_{\text{inf.}} X_i$ na dwóch pozostałych składowych ekorozwoju, tj. społeczeństwie i gospodarce, łącznie w triadzie pojęć stanowiących o istocie ekorozwoju – rysunek 2. Niewątpliwą zaletą przedstawionego algorytmu jest jasny i prosty sposób klastrowania na podstawie $V_{\text{inf.}} X_i$ oraz prezentacja objętości informacji jako metryki w przestrzeni mierzalnej nie tylko w formie metodycznej – tabela 1, ale również w ujęciu geometrycznym, izomorficznym (wzajemnie jednoznacznym) jako „miotełka obiektów" z zapiętym układem współrzędnych w przestrzeni $R^{(nk)+1}$ – rysunek 1, odległość metryzuje objętość informacji.

W zaprezentowanym zagadnieniu dla odległości Euklidesa przedstawione powyżej rozważania można uogólnić na zagadnienia dalsze jako:

$$\begin{bmatrix} d_{O_1} = \|O_1\|_2 \\ d_{O_2} = \|O_2\|_2 \\ d_{O_3} = \|O_3\|_2 \\ \cdots \\ d_{O_n} = \|O_n\|_2 \end{bmatrix} \Rightarrow V_{\text{inf.}} X_i = \prod_1^n \|O_i\|_2$$

Przeprowadzone badania stanowią inspirację dla rozwoju analiz danych, które mogą pogłębić wnioskowanie, zwłaszcza w ujęciu porównawczym rozpatrywanych zagadnień, stanowiąc uzupełnienie analityki dla wielu problemów na gruncie taksonomii z elementami geometrii obliczeniowej.

Literatura

Ackoff, R.L. (1969). *Decyzje optymalne w badaniach stosowanych.* Warszawa: PWN.

Coombs, C.H, Dawes, R.M., Tversky, A. (1977). *Wprowadzenie do psychologii matematycznej.* Warszawa: PWN.

Górka, K. (2007). Wdrażanie koncepcji rozwoju zrównoważonego i trwałego. *Ekonomia i Środowisko*, *2*/32.

Jajuga, K. (1993). *Statystyczna analiza wielowymiarowa*. Warszawa: PWN.

Kaczmarczyk, B. (2015). *Wielowymiarowe ujęcie estymacji wartości rynkowej przedsiębiorstw na przykładzie branży energetycznej* (rozprawa doktorska, materiał niepublikowany). Kraków: Uniwersytet Ekonomiczny.

Popławski, Ł. (2009). *Uwarunkowania ekorozwoju gmin wiejskich na obszarach chronionych województwa świętokrzyskiego.* Warszawa: PWN.

Stevens, S.S. (1946). On the Theory of Scales of Measurement. *Science*, *CIII*, Jun. 7, 2684.

Zaufal, T. (1983). Perspektywy sozologii w ekorozwoju. *Aura*, *3*.

TAXONOMICAL ANALYSIS INFORMATION OF DATA CUBES – VOLUME OF INFORMATION ON THE CHOSEN EXAMPLE OF ECO DEVELOPMENT

Abstract

The term “volume of information”, which was mentioned in the title is a keyword for relatively broad and recent issue of taxonomy and clustering objects or data cubes in regional depiction. The main research in data analysis is focused on variables and development of countless methods, techniques and tools.

The main goal of this article is to present the principles for objects clustering respecting an analysis of vertical data cube and usage of the term “volume of information” in connection with quantitative term of closeness on one side of the objects within the cube, on the other of the possibility of analysis of the metric data cubes in relation to each other. The

object clustering case study for this thesis were used the data of chosen communities from Świętokrzyskie voivodship in connection with their eco development.

Keywords: volume of information, data cube, eco development

Translated by Bogusław Kaczmarczyk

JEL codes: C02, C81,C82