Metoda biplot w interpretacji danych złożonych (CDA) w geologii

Krzysztof Labus[1], Małgorzata Labus[1]

K. Labus

M. Labus

The biplot method in Compositional Data Analysis (CDA) in geology. Prz. Geol., 58: 436–442.

A b s t r a c t. Compositional data consist of compositions of parts summing to some whole (e.g.: 100%, 1). Such "closed" data are very popular in geochemistry, sedimentology, hydrogeology and environmetrics. This paper presents the data visualization method by means of the biplot method — a statistical technique widely applied in Compositional Data Analysis. This method enables a graphical display of observations and variables on the same chart, in a way that approximates their correlation. In a biplot the observations are marked with points, and variables — by rays emanating from the origin. Both their lengths and directions are important to the interpretation. The paper presents two examples of implementation of biplots in analysis of hydrogeological and petrological data.

The first example concerns petrological data set for pores and skeleton grains of different Polish sandstones. The biplot visualization unveiled a possible independence of subcompositions of transitional pores (T) and macropores (R), as well as submacropores (S) and skeleton grains (Re). The use of this method made it also possible to demonstrate the rule according to which the share of transitive pores decreases at the advantage of real macropores from younger to older rocks, what is related to diagenesis processes.

The second example concerns interpretation of hydrological data on chemistry of mine water of the "Rydułtowy" coal mine. In this case the graphical interpretation of biplot revealed large relative variation between HCO_3^--Cl^-, HCO_3^--I^- and HCO_3^--Na^+. Besides, on the same basis, possible independence between the ion couples: HCO_3^--Br^-, Ca^{2+}-Na^+, HCO_3^--Cl^-, Br-Na^+, HCO_3^--Cl^- and Ca^{2+}-K^+ was identified and consequently verified using statistical tests. Additionally, taking into account geometrical features of the relevant biplot, a formula was proposed for defining relationships between shares of ions: Ca^{2+}, Mg^{2+}, Na^+ and Cl^-, SO_4^{2-}, HCO_3^-. These relationships are most apparent for poly-ion groundwater of the active exchange zone, where the ratios of the above mentioned components are similar to each other, despite a slight predominance of SO_4^{2-} and HCO_3^- fractions.

Keywords: *Compositional Data Analysis, biplot, groundwater chemical composition, rock porosity*

Dane złożone niezwykle często występują w zagadnieniach geologicznych, np. w hydrogeologii (procentowe udziały jonów w składzie wód), w petrografii (udziały faz krystalicznych w skale) (tab. 1). Dane te można przedstawić w formie wektorów o nieujemnych elementach x_1,...., x_D, stanowiących pewną całość:

$$x_1 + \ldots + x_D = 1$$

Ponieważ składowe w równaniu sumują się do jedności (100%), nie są zmiennymi niezależnymi, zatem wymagają specyficznego podejścia, które zapewnia zespół procedur określany jako analiza danych złożonych — *Compositional Data Analysis* (CDA), wprowadzona przez Aitchisona w 1986 r. Wizualizacja i interpretacja danych złożonych jest możliwa na diagramach trójkątnych (np. Aitchison, 1986; Labus & Labus, 2006) lub dzięki technice biplot zaproponowanej przez Gabriela (1971).

Wykresy typu biplot i ich interpretacja

Technika biplot pozwala na przedstawienie macierzy danych, czyli zestawień obserwacji (próbek) i opisujących je zmiennych, na tym samym wykresie, w sposób, który opisuje ich wzajemne zależności. Na wykresach typu biplot obserwacje są zazwyczaj zaznaczone jako punkty, natomiast zmienne jako wektory o wspólnym początku.

Obliczenia prowadzące do konstrukcji wykresu rozpoczyna utworzenie macierzy danych, obejmującej w wierszach poszczególne przypadki, a w kolumnach opisujące je zmienne (fragment tego typu macierzy przedstawia tabela 1). Dane początkowe są następnie poddawane przekształceniom, np. centrowaniu, normalizacji lub najczęściej przekształceniom logarytmicznym. Przetworzona macierz jest następnie poddawana dekompozycji na wartości osobliwe (*Singular Value Decomposition*). Operacja ta pozwala na wydobycie wymiarów przekazujących maksimum wariancji zawartej w macierzy początkowej („przestrzeni danych"). Ostatecznym krokiem jest skalowanie uzyskanych wektorów i współrzędnych punktów reprezentujących przypadki.

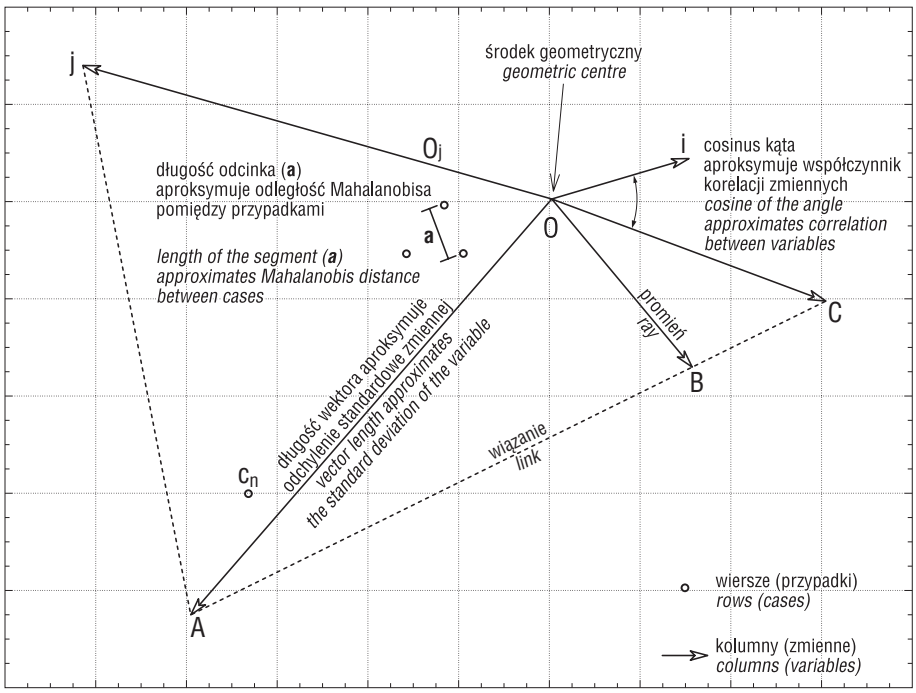
Tab. 1. Przykład zestawu danych złożonych
Table 1. Example of compositional dataset

Obserwacje *Samples*	Zmienne, *Variables* [% mval]						
	Na	K	Mg	Ca	Cl	SO_4	HCO_3
1	40,7	0,4	5,2	3,7	47,8	1,9	0,3
2	34,5	1,0	2,2	12,3	36,7	11,6	1,7
3	10,2	0,9	6,2	32,6	10,7	39,3	0,1

Algorytm obliczeń prowadzących do uzyskania biplotu może wydawać się dość skomplikowany i dla celów praktycznych jest zalecane posługiwanie się odpowiednim oprogramowaniem, np. *XLS-Biplot* (dostępne pod adresem: http://tukey.upf.es/xls-biplot) (Udina, 2005) lub wykorzystanym przez autorów niniejszego artykułu *CoDaPack* (http://ima.udg.edu/~thio/#Compositional%20Data%20Package) (Thió-Henestrosa & Martín-Fernández, 2005).

Środek biplotu *O* reprezentuje środek ciężkości (centroid) zestawu danych; wierzchołki wektorów (*A*, *B*, *C*, *i*, *j*) odpowiadają zmiennym kompozycji, znaczniki (punkty) c_n —

[1]Wydział Górnictwa i Geologii, Politechnika Śląska, ul. Akademicka 2, 44-100 Gliwice; krzysztof.labus@polsl.pl



Ryc. 1. Interpretacja parametrów biplotu (wg Labus, 2005; nieco zmienione)
Fig. 1. Interpretation parameters of a biplot (after Labus, 2005; slightly modified)

oznaczają poszczególne przypadki (ryc. 1). Odcinek łączący punkt *O* z wierzchołkiem *j* jest nazywany promieniem *Oj*, odcinek łączący dwa wierzchołki *i* oraz *j* — wiązaniem *ij*. Zależnie od wyjaśnianego przez biplot zasobu wariancji, na podstawie wiązań i promieni można wnioskować na temat struktury kowariancji zestawu danych.

Długość odcinka pomiędzy znacznikami przypadków (*a*) aproksymuje odległość Mahalanobisa pomiędzy tymi przypadkami. Odległość Mahalonobisa jest odległością między dwoma punktami w przestrzeni *n*-wymiarowej. W przeciwieństwie do odległości euklidesowej ujmuje ona korelację wewnątrz zestawu danych i jest niezależna od efektu skali pomiarów.

Podstawowe cechy biplotu (Aitchison, 2003b; Aitchison & Greenacre, 2002):

Właściwość 1. Długości wektorów, reprezentujące zmienne (kolumny), aproksymują odchylenia standardowe odpowiednich proporcji logarytmicznych (logarytmów proporcji pomiędzy odpowiednimi parami zmiennych; ang. *log-ratio*). Krótkie wiązania pomiędzy punktami zmiennych (końcami wektorów) wskazują, iż proporcje pomiędzy zmiennymi są relatywnie stałe, podczas gdy długie wiązania sugerują wyższe wartości względnej wariancji:

$$|ij|^2 \approx \mathrm{var}\left[\ln\left(x_i/x_j\right)\right], \text{ oraz } |Oi|^2 \approx \mathrm{var}\left\{\ln\left[x_i/g(x)\right]\right\},$$

gdzie $g(x)$ oznacza środek geometryczny.

Właściwość 2. Wartości cosinusa kątów pomiędzy wiązaniami biplotu aproksymują współczynniki korelacji pomiędzy proporcjami logarytmicznymi. Jeżeli wiązania *ij* oraz *kl* przecinają się w punkcie *M*, wówczas:

$$\cos|iMk| = corr\left[\ln\left(x_i/x_j\right), \ln\left(x_k/x_l\right)\right]$$

W przypadku wiązań prostopadłych do siebie $\cos|iMk| \approx 0$, mamy więc do czynienia z brakiem korelacji pomiędzy proporcjami logarytmicznymi. Jest to cecha przydatna podczas poszukiwań niezależności między zmiennymi, możliwej do zweryfikowania na podstawie opisanych dalej testów niezależności (zob. tab. 1 i 2). Zastosowanie testu niezależności winno być poprzedzone potwierdzeniem zgodności rozkładów analizowanych kompozycji z rozkładem logarytmiczno-normalnym za pomocą przedstawionych tu testów.

Właściwość 3. Jeżeli pewien podzbiór wektorów (np. wektory o końcach *j* i *C* — ryc. 1) jest współliniowy, wówczas zmienność związanej z nim subkompozycji jest jednowymiarowa. Jeśli zaś punkty reprezentujące kolumny (zmienne) są usytuowane wzdłuż linii prostej (np. *A*, *B*, *C* — ryc. 1), to model opisujący taką zależność może zostać wyprowadzony na podstawie względnych długości ich wiązań. Oznacza to, iż jeśli dane trzy zmienne *A*, *B* i *C* leżą na linii prostej, a odległości *AB* i *BC* wynoszą odpowiednio α i β, to prawdziwe jest równanie (ang. *log-contrast*):

$$\beta \ln(A) + \alpha \ln(C) - (\alpha+\beta) \ln(B) = \text{const, czyli } (A/B)^{\beta} \approx (B/C)^{\alpha}$$

Testowanie hipotez o zgodności danego rozkładu z rozkładem normalnym

Porównanie obliczonych wartości odpowiednich statystyk z wartościami krytycznymi pozwala na podjęcie decyzji o przyjęciu lub odrzuceniu badanej hipotezy o normalności danego rozkładu lub rozkładów. Wyższe wartości statystyk odpowiadają niższemu poziomowi istotności. Sekwencje wartości z_i (obliczane dla testów statystyk rozkładów brzegowych oraz dwuwymiarowych), uszeregowane w porządku rosnącym, są używane w wyrażeniu Q_A w teście Andersona–Darlinga, Q_C w teście Cramera–von Misesa oraz Q_W w teście Watsona (Pawlowsky-Glahn & Buccianti, 2002; Aitchison, 2003a) (tab. 2 i 3).

Wartości proporcji logarytmicznych (*log-ratio*) w rozkładach brzegowych można przedstawić jako:

$$y_{ri} = \ln\left(x_{ri}/x_{rN}\right), \text{ gdzie } r = 1, \ldots, N.$$

Tab. 2. Formuły i wartości krytyczne testów statystyk rozkładów brzegowych (Labus, 2005)
Table 2. Formulas and critical values of marginal tests statistics (Labus, 2005)

Testy, *Tests*	**Poziom istotności,** *Significance level* **[%]**			
	10	**5**	**2,5**	**1**
Andersona–Darlinga, *Anderson–Darling* $Q_A = \left\{-\frac{1}{N}\sum_{i=1}^{N}(2i-1)\left[\ln z_i + \ln(1-z_{N+1-i})\right] - N\right\}\left(1+\frac{4}{N}-\frac{25}{N^2}\right)$	0,656	0,787	0,918	1,092
Cramera–von Misesa, *Cramer–von Mises* $Q_C = \left[\sum_{i=1}^{N}\left(z_i - \frac{2i-1}{2N}\right)^2 + \frac{1}{12N}\right]\left(1+\frac{1}{2N}\right)$	0,104	0,126	0,148	0,178
Watsona, *Watson* $Q_W = Q_C - N\left(\bar{z}-\frac{1}{2}\right)^2\left(1+\frac{1}{2N}\right)$, gdzie $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$	0,096	0,116	0,136	0,136

Tab. 3. Formuły i wartości krytyczne testów statystyk rozkładów dwuwymiarowych (Labus, 2005)
Table 3. Formulas and critical values of bivariate angle test statistics (Labus, 2005)

Testy, *Tests*	**Poziom istotności,** *Significance level* **[%]**			
	10	**5**	**2,5**	**1**
Andersona–Darlinga, *Anderson–Darling* $Q_A = -\frac{1}{N}\sum_{i=1}^{N}(2i-1)\left[\ln z_i + \ln(1-z_{N+1-i})\right] - N$	1,933	2,492	3,070	3,857
Cramera–von Misesa, *Cramer–von Mises* $Q_C = \left\{\left[\sum_{i=1}^{N}\left(z_i - \frac{2i-1}{2N}\right)^2 + \frac{1}{12N}\right] - \frac{0,4}{N} + \frac{0,6}{N^2}\right\}\left(1+\frac{1}{N}\right)$	0,347	0,461	0,581	0,743
Watsona, *Watson* $Q_W = \left[\sum_{i=1}^{N}\left(z_i - \frac{2i-1}{2N}\right)^2 - N\left(\bar{z}-\frac{1}{2}\right)^2 - \frac{0,2}{12N} + \frac{0,1}{N^2}\right]\left(\frac{N+0,8}{N}\right)$	0,152	0,187	0,221	0,267

Wartość z_i jest obliczana jako wartość funkcji skumulowanego rozkładu normalnego — przy średniej równej 0 i jednostkowym odchyleniu standardowym $N \in (0; 1)$ — na podstawie formuły:

$$\Phi\left(\left(y_{i1} - \bar{y}_1\right)/s_1\right) = z_{i1},$$

gdzie:
s — odchylenie standardowe,
Φ — funkcja skumulowanego rozkładu normalnego.

Wartości z_i zastosowane w testach statystyk rozkładów dwuwymiarowych są obliczane jako:

$$z_i = \theta_i/(2\pi),$$

gdzie:

$$\theta_1 = \arctan\left(u_{i2}/u_{i1}\right) + 0,5\left[1-\operatorname{sgn}(u_{i1})\right]\pi + 0,5\left[1+\operatorname{sgn}(u_{i1})\right]\left[1-\operatorname{sgn}(u_{i2})\right]\pi$$

oraz

$$u_{i1} = \frac{\left(y_{i1} - \bar{y}_1\right)s_2}{\sqrt{s_1^2 s_2^2 - s_{12}^2}} - \frac{\left(y_{i2} - \bar{y}_2\right)s_{12}}{s_2\sqrt{s_1^2 s_2^2 - s_{12}^2}};$$

$$u_{i2} = \frac{\left(y_{i2} - \bar{y}_2\right)}{s_2}$$

Funkcja signum (sgn) jest zdefiniowana jako:

$$\operatorname{sgn}(x) = \begin{cases} -1, & \text{dla } x < 0, \\ 0, & \text{dla } x = 0, \\ -1, & \text{dla } x > 0 \end{cases}$$

Test niezależności subkompozycyjnej

Relacje pomiędzy kompozycjami, których niezależność wskazuje graficzna interpretacja biplotu, należy zbadać za pomocą testu niezależności subkompozycyjnej (*subcompositional independence test*; Aitchison, 2003b). Test ten polega na porównaniu wartości wyrażenia:

$$N\left[\ln\left|\hat{\Sigma}_{11}\right|\left|\hat{\Sigma}_{22}\right| - \ln\begin{vmatrix}\hat{\Sigma}_{11} & \hat{\Sigma}_{12}\\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22}\end{vmatrix}\right]$$

wobec wyższych kwantyli zmiennej o rozkładzie:

$$\chi^2\left[(c-1)(d-c)\right],$$

gdzie:
$\hat{\Sigma}_{ij}$ — macierze kowariancji z próby,
d — liczba wymiarów analizowanej kompozycji (dla kompozycji złożonej z D elementów $d = D - 1$),
c — liczba wymiarów analizowanej subkompozycji (dla subkompozycji złożonej z C elementów $c = C$).

Przykłady wizualizacji danych złożonych metodą biplot

W artykule zaprezentowano dwa przykłady zastosowania interpretacji wykresów biplot dla danych złożonych — petrologicznych i hydrogeologicznych.

Przykład 1. Pierwszy z przykładów przedstawia wizualizację wyników badań porozymetrycznych piaskowców, obejmujących procentowe udziały porów (tab. 4) oraz szkieletu ziarnowego. Próbki wykorzystane do badań reprezentowały różnowiekowe piaskowce pochodzące z terenu Polski (Labus & Labus, 2006). Oprócz uzyskania możliwie pełnej charakterystyki przestrzeni porowej tych skał zaplanowano analizę zależności udziałów porów od ich genezy i pozycji stratygraficznej. Przedstawienie procentowego udziału porów w próbkach na wykresach typu histogramów nie daje wystarczającego obrazu do porównania rozkładu porów w poszczególnych grupach skał (ryc. 2). Wygląd wykresu zależy m.in. od kolejności próbek, nie pozwala ponadto na wyciąganie wniosków dotyczących powiązań pomiędzy poszczególnymi grupami porów. Dlatego też podjęto próbę przedstawienia uzyskanych danych na wykresie typu biplot. Tym razem 100% całość tworzą nie tylko same pory w skale, ale pory wraz ze szkieletem ziarnowym (oznaczonym na wykresie symbolem Re) (ryc. 3).

Określenie „skumulowana wariancja" (po lewej stronie, u góry diagramu) odnosi się do wariancji wyjaśnionej przez $D - 1$ składowych głównych (dla D analizowanych zmiennych), przy założeniu, iż wszystkie ze składowych wyjaśniają w sumie 100% wariancji. W analizowanym przykładzie pierwsza ze składowych wyjaśnia 80% wariancji, podczas gdy druga 11%. Obydwie pozwalają na interpretację 91% wariancji w obrębie analizowanego zestawu danych, co odpowiada równocześnie wariancji wyjaśnionej przez biplot — narzędzie dwuwymiarowego obrazowania zależności w obrębie populacji.

Tab. 4. Graniczne średnice wyróżnionych grup porów
Table 4. Group ranges of distinguished pore diameters

Klasy porów *Pore classes*	**Symbol**	**Średnica [m]** *Diameter [m]*
Pory przejściowe *Transitional pores*	T	10^{-8}–10^{-7}
Submakropory *Submacropores*	S	10^{-7}–10^{-6}
Makropory właściwe *Real macropores*	R	10^{-6}–10^{-4}
Pory nadkapilarne *Over capillary pores*	O	$>10^{-4}$

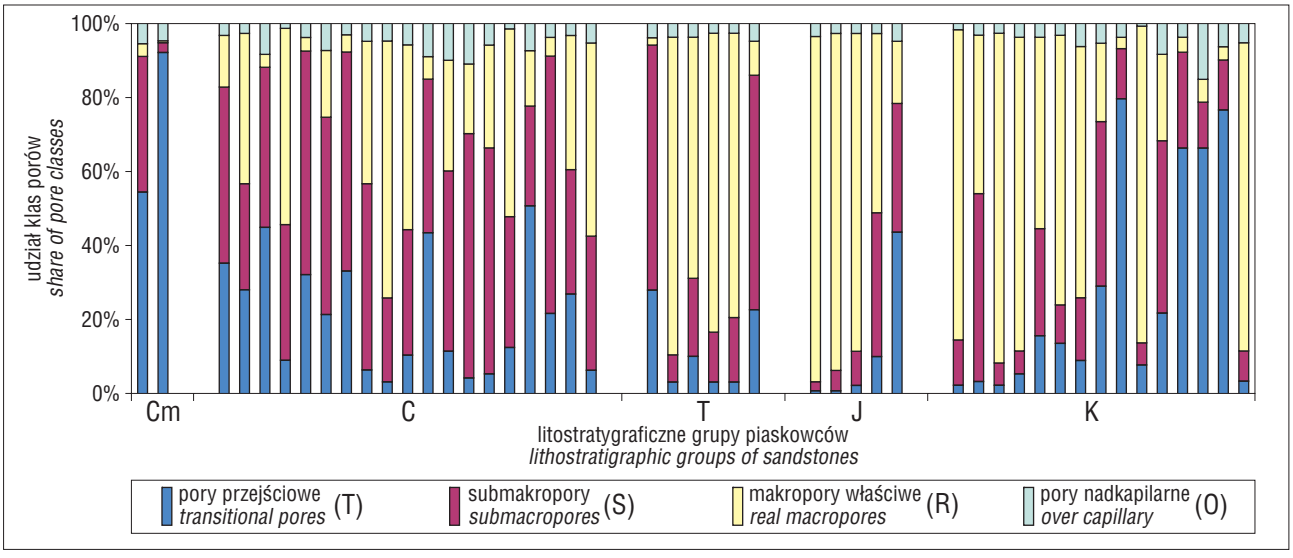
Kolejny diagram (ryc. 4) przedstawia środki geometryczne wyróżnionych wcześniej grup piaskowców na tle promieni biplotu. Usytuowanie punktów świadczy, iż udział porów przejściowych (T) spada na rzecz makroporów (R), porządkując zespoły próbek w następującej kolejności: J ~> K > T > C > K-flysch. Spadek udziału najmniejszych porów — przejściowych (T) może być związany ze zmniejszającym się zaawansowaniem diagenezy, zależnym od wieku piaskowców. Skały starsze powinny zawierać mniej porów o dużej objętości.

Wstępna interpretacja biplotu danych petrologicznych (ryc. 3) mogłaby być następująca:

1) Najdłuższymi wiązaniem biplotu jest T–R, co oznacza, iż najwyższą zmiennością cechują się relacje między udziałami porów przejściowych i makroporów.

2) Wiązania S–Re oraz T–R są do siebie niemal prostopadłe, co sugeruje potencjalną niezależność odpowiednich proporcji logarytmicznych udziałów submakroporów i szkieletu ziarnowego oraz porów przejściowych i makroporów.

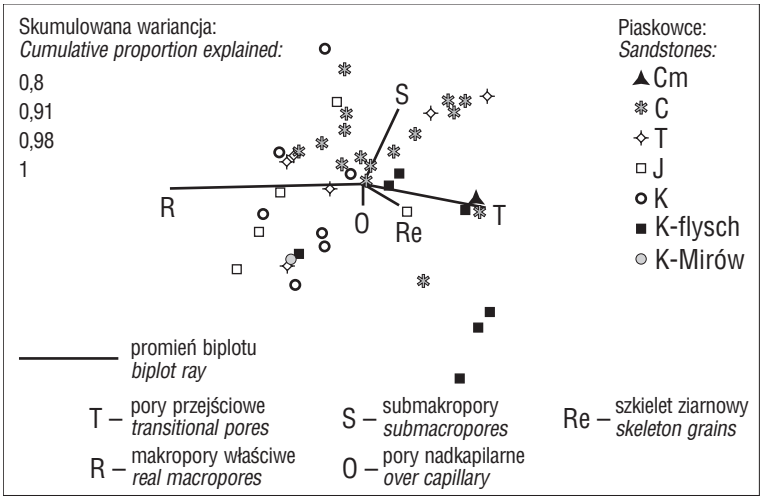
Przybliżona współliniowość wierzchołków R, T, Re wskazuje na prawdopodobieństwo jednowymiarowej



Ryc. 2. Diagramy udziałów klas porów w analizowanych piaskowcach; Cm — kambryjski piaskowiec kwarcytowy (Góry Świętokrzyskie), C — piaskowce górnokarbońskie Górnośląskiego Zagłębia Węglowego, T — piaskowce triasowe (Góry Świętokrzyskie), J — piaskowce jurajskie (Góry Świętokrzyskie), K — piaskowce kredowe (łącznie)
Fig. 2. Bar charts of pore classes in the analysed groups of sandstones; Cm — Cambrian quartzitic sandstone (Holy Cross Mts.), C — Upper Carboniferous sandstones (Upper Silesian Coal Basin), T — Triassic sandstones (Holy Cross Mts.), J — Jurassic sandstones (Holy Cross Mts.), K — Cretaceous sandstones (all types)

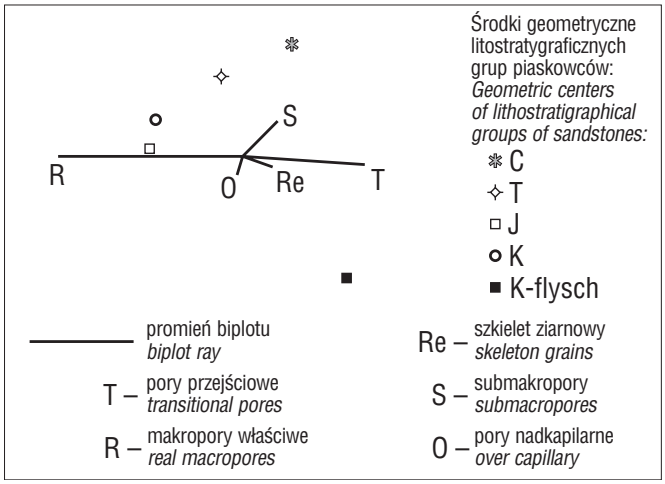
zmienności reprezentowanej przez nie subkompozycji, zatem *log-contrast*: $\beta \log (T) + \alpha \ln (R) - (\alpha + \beta) \ln (Re) = \text{const}$, ma postać: $2{,}71 \ln (T) + \ln (R) - 3{,}71 \ln (Re) = \text{const}$, czyli w przybliżeniu $(T^3 R)/Re^4 = \text{const}$. Tak przedstawiona zależność zmusza do odpowiedzi na pytanie, czy zaprezentowana formuła ma jakikolwiek sens geologiczny. Z petrologicznego punktu widzenia nie odpowiada ona znanym regułom, nie oznacza to jednak, iż jest nieprzydatna. Godna weryfikacji byłaby jej potencjalna wartość jako charakterystyki (swoistego kodu) odmiennych skał klastycznych — różniących się wiekiem, litologią, stopniem diagenezy etc.

Do zbadania (zidentyfikowanych na podstawie Właściwości 2) niezależności proporcji logarytmicznych udziałów submakroporów i szkieletu ziarnowego oraz porów przejściowych i makroporów zastosowano tzw. test niezależności subkompozycyjnej (Aitchison, 2003b). Zastosowanie tego testu wymagało uprzedniej weryfikacji zgodności rozkładów analizowanych proporcji z rozkładem logistyczno-normalnym. Weryfikacji dokonano za pomocą testów Andersona–Darlinga, Cramera–von Misesa oraz Watsona. Rezultaty testów statystyk, których formuły i wartości krytyczne przedstawiono w tabeli 1 i 2, ilustruje tabela 5.



Ryc. 3. Biplot subpopulacji rozkładu porów i szkieletu ziarnowego; Cm — kambryjski piaskowiec kwarcytowy (Góry Świętokrzyskie), C — piaskowce górnokarbońskie Górnośląskiego Zagłębia Węglowego, T — piaskowce triasowe (Góry Świętokrzyskie), J — piaskowce jurajskie (Góry Świętokrzyskie), K — piaskowce kredowe z Dolnego Śląska, K-flysch — górnokredowe piaskowce fliszowe Beskidu Śląskiego, K-Mirów — piaskowiec kredowy z okolic Mirowa (Wyżyna Śląsko-Krakowska)

Fig. 3. Biplot of pores and skeleton grains; Cm — Cambrian quartzitic sandstone (Holy Cross Mts.), C — upper Carboniferous sandstones (Upper Silesian Coal Basin), T — Triassic sandstones (Holy Cross Mts.), J — Jurassic sandstones (Holy Cross Mts.), K — Cretaceous sandstones (Lower Silesia), K-flysch — upper Cretaceous flysch sandstones (Beskid Śląski region), K-Mirów — Cretaceous sandstone from Mirów area (Silesian-Cracow Upland)



←

Ryc. 4. Biplot rozkładu porów i szkieletu ziarnowego z zaznaczeniem środków geometrycznych litostratygraficznych grup piaskowców (zwraca uwagę nieznaczna reorientacja promieni biplotu w porównaniu do ryc. 3)

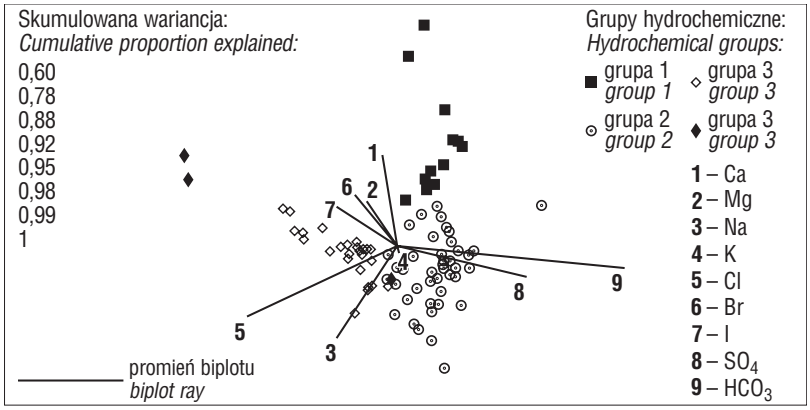
Fig. 4. Biplot of distribution of pores and skeleton grains and geometrical centers of lithostratigraphical groups of sandstones (note a slight reorientation of biplot rays in comparison to that from Fig. 3)

Tab. 5. Rezultaty testów zgodności z rozkładem normalnym oraz testu niezależności rozkładów ln (T/R) i ln (S/Re) — dla przykładu 1

Table 5. Multivariate normality tests and compositional independence test of ln (T/R) and ln (S/Re) for the example 1

	Anderson–Darling	Cramer–von Mises	Watson
ln (T/R) — rozkład brzegowy *marginal distribution*	0,776	0,116	0,114
ln (S/Re) — rozkład brzegowy *marginal distribution*	0,709	0,115	0,099
Rozkład dwuwymiarowy *Bivariate distribution*	0,862	0,170	0,089
ln (T/R) – ln(S/Re) — test niezależności *independence test*	–	–	0,776

Wyniki obliczeń dowodzą, iż analizowane rozkłady proporcji nie wykazują rozbieżności z rozkładem normalnym na poziomie istotności około 5% wszystkich przeprowadzonych testów. Test niezależności subkompozycyjnej wykazuje niezależność pomiędzy ln (T/R) – ln (S/Re) z prawdopodobieństwem 0,776, co równocześnie oznacza niezależność subkompozycji (T, R) od (S, Re) — spostrzeżenie uczynione dzięki wstępnej interpretacji graficznej biplotu. Równocześnie uzyskano potwierdzenie braku korelacji pomiędzy proporcjami logarytmicznymi ln (T/R) i ln (S/Re).



←

Ryc. 5. Biplot danych składu wód kopalnianych KWK *Rydułtowy* (grupy hydrochemiczne)

Fig. 5. Biplot of mine waters chemistry — the *Rydułtowy* coal mine (hydrochemical groups)

Przykład 2. W drugim z przykładów (ryc. 5) zobrazowano kompozycje obejmujące udziały procentowe jonów (% mval) głównych oraz jonu jodkowego i bromkowego w próbkach wód kopalnianych KWK *Rydułtowy*, zaliczonych do odmiennych grup hydrochemicznych (Labus, 2007) (tab. 6). W tym przypadku pierwsza ze składowych wyjaśnia 60% wariancji, podczas gdy druga 18%, co łącznie odpowiada 78% wariancji wyjaśnianej przez biplot. W analizowanym przykładzie wariancję związaną z poszczególnymi składowymi można interpretować jako efekt działania procesów lub zjawisk formujących skład wód. Wnioskowanie oparte na składowych wyjaśniających wysoki zasób wariancji (zwykle powyżej 10%) jest obarczone najmniejszą niepewnością.

Interpretacja biplotu danych hydrogeologicznych (ryc. 5) pozwala na wyciągnięcie wniosków:

1) Najdłuższymi wiązaniami biplotu są: HCO_3–Cl (9–5 na ryc. 5), HCO_3–I (9–7) oraz HCO_3–Na (9–3), co sugeruje, iż najwyższą zmiennością cechują się relacje między udziałami tych właśnie składników (w badanych próbkach brak stałej proporcji pomiędzy nimi). W całej analizowanej populacji jony HCO_3^- pochodzą z innego źródła niż wymienione składniki wód. Zróżnicowanie wewnątrzgrupowe jest mniejsze, co uwidacznia odrębne, skupione położenie punktów reprezentujących poszczególne grupy względem osi tworzonej przez wiązanie, np. HCO_3–Cl (9–5).

2) Wiązania HCO_3–Br (9–6) oraz Ca–Na (1–2) są do siebie prawie prostopadłe, podobnie jak HCO_3–Cl (9–5) oraz Br–Na (6–3) i HCO_3–Cl oraz Ca–K (1–4), co sugeruje niezależność odpowiednich par elementów. Równocześnie zmienne połączone wiązaniami są ze sobą ujemnie skorelowane (np. wzrost udziału HCO_3^- jest sprzężony ze spadkiem Br). Generalny wzrost udziałów Br i spadek Na, na rzecz udziałów Ca, mogą wskazywać na obecność w badanej populacji wód w różnym stopniu zmienionych w procesie odwrotnej wymiany jonowej: Na zamiast Ca.

3) Przybliżona współliniowość wierzchołków Ca, Mg, Na oraz Cl, SO_4, HCO_3 sugeruje jednowymiarową zmienność odpowiadających im subkompozycji. Pozwala to na określenie formuły zależności (*log-contrast*) pomiędzy udziałami wymienionych jonów:

$$3 \ln(HCO_3) + \ln(Cl) - 4 \ln(SO_4) = -3{,}4$$

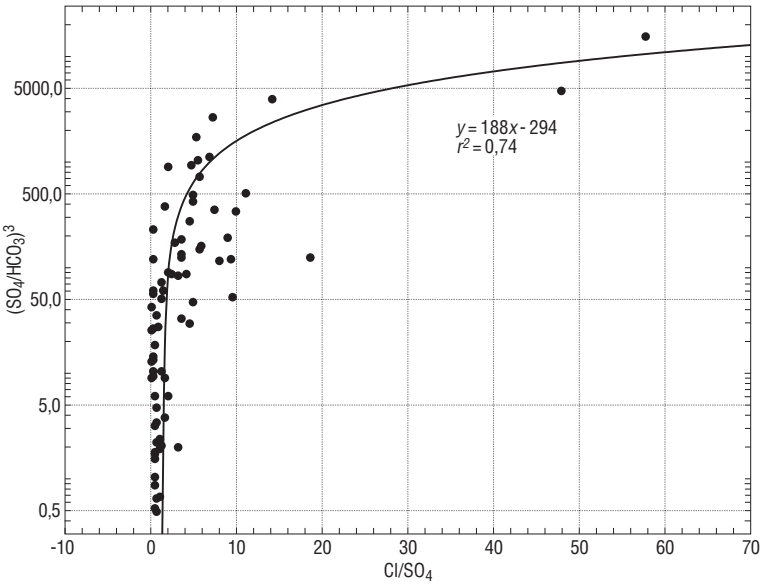
czyli:

$$(Cl/SO_4) \sim (SO_4/HCO_3)^3$$

Zależność ta jest najbardziej wyraźna dla wód wielojonowych strefy aktywnej wymiany, gdzie proporcje wymienionych składników są do siebie zbliżone, przy niewielkiej przewadze udziałów SO_4^{2-} i HCO_3^- (ryc. 6).

Tab. 6. Hydrochemiczne grupy wód kopalnianych KWK *Rydułtow*y
Table 6. Hydrochemical mine waters groups — the *Rydułtowy* coal mine

Grupa *Group*	**Wzór Kurłowa** *Hydrochemical characteristics (Kurlov transcription)*
1	$M^{0,31-2,24} \frac{Cl^{2,5-38,2} SO_4^{28,0-87,3} HCO_3^{8,0-59,6}}{Na^{5,3-50,6} Ca^{33,3-70,6} Mg^{11,9-50,6} K^{0,7-6,5}} (I^{0,0-0,6} Fe^{0,0-9,5})$
2	$M^{0,24-9,71} \frac{Cl^{1,6-52,2} SO_4^{22,1-77,5} HCO_3^{9,7-70,5}}{Na^{46,5-65,0} Ca^{0,6-38,2} Mg^{2,1-18,6} K^{0,8-9,5}} (I^{0,0-1,6} Fe^{0,0-3,2})$
3	$M^{2,69-148,50} \frac{Cl^{53,5-98,2} SO_4^{0,7-37,5} HCO_3^{0,0-21,6}}{Na^{62,8-94,5} Ca^{0,5-19,4} Mg^{2,0-16,7} K^{0,0-4,8}} (I^{0,4-30,9} Fe^{0,0-3,5})$
4	$M^{215,1-217,7} \frac{Cl^{99,6-99,9} SO_4^{0,0-0,4} HCO_3^{0,0}}{Na^{78,8-79,4} Ca^{10,1-11,5} Mg^{9,8} K^{0,4-0,8}} (I^{19,0-72,7} Fe^{6,0-9,5})$



Ryc. 6. Zależność między (Cl/SO_4) i $(SO_4/HCO_3)^3$ w wodach kopalnianych KWK *Rydułtowy*

Fig. 6. The relationship of (Cl/SO_4) vs. $(SO_4/HCO_3)^3$ — the *Rydułtowy* mine

Do zbadania (zidentyfikowanych na podstawie Właściwości 2) niezależności proporcji logarytmicznych udziałów ln (HCO_3/Cl) i ln (Ca/K) zastosowano, podobnie jak w przykładzie 1, test niezależności subkompozycyjnej. Weryfikacji dokonano również za pomocą wymienionych w tabelach 1 i 2 testów. Rezultaty odpowiednich testów statystyk przedstawia tabela 7.

Tab. 7. Rezultaty testów zgodności z rozkładem normalnym oraz testu niezależności rozkładów proporcji wybranych subkompozycji — dla przykładu 2
Table 7. Multivariate normality tests and compositional independence test of selected subcompositions for the example 2

	Anderson–Darling	Cramer–von Mises	Watson
ln (HCO_3/Cl) — rozkład brzegowy *marginal distribution*	1,089	0,174	0,159
ln (Ca/K) — rozkład brzegowy *marginal distribution*	1,010	0,169	0,140
Rozkład dwuwymiarowy *Bivariate distribution*	0,677	0,136	0,130
ln (HCO_3/Cl) – ln (Br/K) — test niezależności *independence test*	–	–	0,999

Przeprowadzone obliczenia dowodzą, iż kompozycje ln (HCO_3/Cl) oraz ln (Ca/K) nie wykazują rozbieżności z rozkładem normalnym na poziomie istotności około 1% dla testów statystyk brzegowych oraz na poziomie powyżej 10% dla testów zgodności z rozkładem dwuwymiarowym. Obliczone za pomocą testu niezależności subkompozycyjnej prawdopodobieństwo niezależności wymienionych subkompozycji wynosi ponad 0,99. Oznacza to, iż niejednakowe są zjawiska regulujące związki pomiędzy odpowiednimi parami zmiennych (udziałów składników wód). Relacja HCO_3 z Cl jest najprawdopodobniej regulowana zjawiskami rozcieńczania wód reliktowych przez infiltracyjne. Wzbogacenie wód w Ca jest zjawiskiem konkurencyjnym w stosunku do przyrostu K i skorelowanego z nim udziału Na. Wraz ze wzrastającą rolą wód reliktowych w wodach kopalnianych rosną proporcje Ca/Na, a tym samym Ca/K. Niezależność obserwowanych zjawisk pozwala na wniosek, iż proces rozpuszczania $CaCO_3$, prowadzący do równoczesnego wzrostu udziałów Ca i HCO_3, ma niewielkie znaczenie podczas formowania chemizmu większości wód kopalnianych KWK *Rydułtowy*. Wyjątek stanowią jedynie wody grupy 1 oraz część wód grupy 2, których punkty projekcyjne zawierają się pomiędzy ramionami biplotu odpowiadającymi zmiennym Ca i HCO_3.

Podsumowanie

Zaprezentowana w artykule, w zarysie, metoda interpretacji i wizualizacji danych, jaką jest biplot, pozwala na przedstawienie obserwacji i zmiennych na tym samym wykresie, w sposób opisujący ich wzajemne zależności. Wstępna interpretacja biplotu danych złożonych pozwala m.in. na ocenę względnego zróżnicowania zmiennych (poprzez oszacowanie względnych wartości ich wariancji) i zależności korelacyjnych pomiędzy zmiennymi. Spostrzeżenia poczynione na podstawie diagramu można zweryfikować za pomocą obliczeń statystycznych, pośród których istotne miejsce zajmują testy hipotez o zgodności rozkładów analizowanych par zmiennych z rozkładem normalnym. Metoda biplot ułatwia interpretację złożonych danych geologicznych, charakteryzujących się zarówno wysoką liczbą obserwacji, jak i badanych cech.

Literatura

AITCHISON J. 1986 — The Statistical Analysis of Compositional Data, Monographs on Statistics and Applied Probability. Chapman and Hall, London.
AITCHISON J. 2003a — The Statistical Analysis of Compositional Data. Blackburn Press, New Jersey.
AITCHISON J. 2003b — A Concise Guide to Compositional Data Analysis. CDA Workshop, Girona.
AITCHISON J. & GREENACRE M. 2002 — Biplots of Compositional Data. Appl. Statist.,51: 375– 382.
GABRIEL K.R. 1971 — The biplot display of matrices with application to principal components analysis. Biometrika, 58: 453–467.
LABUS K. 2007 — Identyfikacja procesów formujących chemizm wód podziemnych w warunkach drenażu górniczego, w południowo-zachodniej części Górnośląskiego Zagłębia Węglowego. Zesz. Nauk. PŚl. Gór., 281: 1–247.
LABUS M. 2005 — Compositional data analysis as a tool for interpretation of rock porosity parameters. Geol. Quart., 49, 3: 347–354.
LABUS K. & LABUS M. 2006 — Zastosowanie analizy danych złożonych (CDA) w geologii. Gosp. Sur. Min., 22, 2: 39–52.
PAWLOWSKY-GLAHN V. & BUCCIANTI A. 2002 — Visualization and modeling of sub-populations of compositional data: statistical methods illustrated by means of geochemical data from fumarolic fluids. Int. J. Earth Sci., 91: 357–368
THIÓ-HENESTROSA S. & MARTÍN-FERNÁNDEZ J.A. 2005 — Dealing with compositional data: the freeware CoDaPack. Math. Geol., 37, 7: 773–793.
UDINA F. 2005 — Interactive Biplot construction. J. Stat. Softw., 13, 3: 1–16; http://www.jstatsoft.org/.

Praca wpłynęła do redakcji 10.02.2009 r.
Po recenzji akceptowano do druku 28.12.2009 r.