NAFTA-GAZ

Nafta-Gaz 2023, nr 9, s. 576–583, DOI: 10.18668/NG.2023.09.02

# Wykorzystanie języka R do statystycznej analizy oraz analizy skupień dla danych geochemicznych

## Use of R programming language for statistical analysis and cluster analysis of geochemical data

Marek Janiga

*Instytut Nafty i Gazu – Państwowy Instytut Badawczy*

STRESZCZENIE: W zagadnieniach geologii naftowej metody statystyczne są szeroko stosowane w petrografii, petrofizyce, geochemii, geomechanice, geofizyce wiertniczej czy sejsmice, a analiza skupień jest istotna w klasyfikacji skał – wyznaczaniu stref o pewnych własnościach, np. macierzystych lub zbiornikowych. Artykuł prezentuje użycie metod statystycznych, w tym metod analizy skupień, w procesach przetwarzania i analizy dużych zbiorów różnorodnych danych geochemicznych. Do analiz statystycznych wykorzystano literaturowe dane z analiz składu chemicznego i izotopowego gazów ziemnych. Wyniki zawierały skład chemiczny gazów ziemnych oraz skład izotopowy. Zastosowano algorytmy tzw. nienadzorowanego uczenia maszynowego do przeprowadzenia analizy skupień. Grupowania było przeprowadzone dwiema metodami: k-średnich oraz hierarchiczną. Do zobrazowania wyników grupowania metodą k-średnich można wykorzystać dwuwymiarowy wykres (funkcja *fviz_cluster* języka R). Wymiary na wykresie to efekt analizy głównych składowych (PCA) i są one liniową kombinacją cech (kolumn w tabeli). Wynikiem grupowania metodą hierarchiczną jest wykres nazywany dendrogramem. W artykule dodatkowo zaprezentowano wykresy pudełkowe i histogramy oraz macierz korelacji zawierającą współczynniki korelacji Pearsona. Wszystkie prace wykonano z użyciem języka programowania R. Język R, z wykorzystaniem programu RStudio, jest bardzo wygodnym i szybkim narzędziem do statystycznej analizy danych. Przy użyciu tego języka uzyskanie wymienionych powyżej wykresów, tabeli i danych jest szybkie i stosunkowo łatwe. Wyniki analiz składu gazu wydają się mało zróżnicowane. Mimo to dzięki algorytmom k-średnich i hierarchicznym możliwe było pogrupowanie danych geochemicznych na wyraźnie rozdzielne zespoły. Zarówno wartości składu izotopowego, jak i skład chemiczny pozwalają wyznaczyć grupy, które w inny sposób nie byłyby dostrzegalne.

Słowa kluczowe: analiza skupień, metoda k-średnich, metoda hierarchiczna, skład gazu ziemnego.

ABSTRACT: In petroleum geology, statistical methods are widely used in petrography, petrophysics, geochemistry, geomechanics, well log analysis and seismics, and cluster analysis is important for rock classification – determination of zones with certain properties, e.g., source or reservoir. This paper presents the use of the R language for statistical analysis, including cluster analysis, of large sets of diverse geochemical data. Literature data from analyses of chemical and isotopic composition of natural gases were used for statistical analyses. The results included the chemical composition of the natural gases and the isotopic composition. So-called unsupervised machine learning algorithms were used to perform the cluster analysis. Clustering was performed using two methods: k-means and hierarchical. A two-dimensional graph (function fviz_cluster) can be used to illustrate the results of the k-means clustering. The dimensions in the graph are the result of principal component analysis (PCA) and are a linear combination of the features (columns in the table). The result of hierarchical clustering is a graph called a dendrogram. The paper additionally presents box plots and histograms as well as a correlation matrix containing Pearson correlation coefficients. All work was completed using the programming language R. The R language, using the RStudio software, is a very convenient and fast tool for statistical data analysis. Obtaining the above-mentioned graphs, tables and data is quick and relatively easy, using the R language. The results of the analyses of the composition of the gas appear to have little variation. Nevertheless, thanks to k-means and hierarchical algorithms, it was possible to group the geochemical data into clearly separable groups. Both the isotopic composition values and the chemical composition make it possible to delineate groups that would not otherwise be noticeable.

Key words: cluster analysis, k-means method, hierarchical method, natural gas composition.

Autor do korespondencji: M. Janiga, e-mail: *marek.janiga@inig.pl*

Artykuł nadesłano do Redakcji: 22.02.2023 r. Zatwierdzono do druku: 04.09.2023 r.

## Wstęp

Metody statystyczne są szeroko stosowane w badaniach naukowych, a większość renomowanych czasopism naukowych nie opublikuje artykułu bez rozdziału dotyczącego analizy statystycznej. W naukach geologicznych sytuacja nie jest odmienna. W zagadnieniach geologii naftowej metody statystyczne są szeroko wykorzystywane w petrografii, petrofizyce, geochemii, geomechanice, geofizyce wiertniczej czy sejsmice (Topór, 2021). Analiza skupień jest istotna w klasyfikacji skał – wyznaczaniu stref o pewnych własnościach, np. macierzystych lub zbiornikowych (Topór, 2020).

Język programowania R powstał z myślą o analizie danych. Jego podstawą jest język S, którego pierwsza wersja powstała w 1991 roku i miała służyć studentom Uniwersytetu w Auckland jako pomoc w nauce statystyki. Do pisania programów w języku R można użyć edytora tekstu lub programów nazywanych zintegrowanym środowiskiem programistycznym (IDE). Takim programem jest RStudio, które ułatwia używanie języka R.

RStudio pozwala na szybki i łatwy import danych z plików. W programie R tabela danych jest również nazywana ramką danych. Jednowymiarowa ramka danych (np. nazwy cech) to wektor danych. Dane geochemiczne zostały zaimportowane do programu i stworzyły ramkę danych o nazwie „dane". Ramka ta jest dostępna w katalogu roboczym i nie jest już konieczne podawanie ścieżki do lokalizacji pliku. Dodatkowo zaimportowano dwa zestawy danych: osobno wyniki analiz składu chemicznego („danechem") oraz wyniki analiz składu izotopowego („daneizo").

Programowanie w języku R opiera się na korzystaniu z już zdefiniowanych funkcji. Funkcji podaje się argument (dane) oraz można wyspecyfikować warunki działania danej funkcji. W funkcjach tabela danych jest umieszczana w nawiasach okrągłych, po nazwie tabeli można dodatkowo określić kolumnę, wykorzystując nawiasy kwadratowe, np. *sd*(dane[['CH4']]) lub *sd*(dane$CH4).

## Wykorzystane dane geochemiczne – skład chemiczny i izotopowy

Do analiz statystycznych wykorzystano literaturowe dane z analiz składu chemicznego i izotopowego gazów ziemnych ze złóż miocenu autochtonicznego zapadliska przedkarpackiego (Kotarba, 1992, 1998, 2011; Kotarba i Jawor, 1993; Kotarba et al., 2005; Kotarba i Nagao, 2008). Wyniki zawierały skład chemiczny gazów ziemnych (m.in.: udział węglowodorów, azotu, ditlenku węgla, helu, wodoru, związków siarki) oraz skład izotopowy (m.in.: $\delta^{13}C\text{-}C_1$, $\delta^{13}C\text{-}C_2$, $\delta^{13}C\text{-}C_3$ i $\delta D\text{-}C_1$). Gazy nie są mocno zróżnicowane i raczej nie wyróżnia się wśród nich grup o podobnych cechach.

## Badania eksploracyjne

Ojcem badań eksploracyjnych oraz samego terminu jest John Tukey. W artykule z 1962 roku zaproponował on nowe podejście do statystyki. Miało ono polegać na łatwym zobrazowaniu własności zestawu danych. W kolejnym artykule, z 1972 roku, o tytule precyzyjnie oddającym zawartość publikacji: *Exploratory data analysis*, zaproponował wykresy (np. pudełkowy) i statystyki, które miały temu celowi służyć (Tukey, 1962, 1977; Bruce et al., 2021).

### *Miary położenia i rozproszenia*

Podstawowym zadaniem w badaniach eksploracyjnych jest poznanie „typowych wartości" dla zmiennych (cech), określenie lokalizacji większości analizowanych danych. Można do tego wykorzystać wartości takie jak: średnia, średnia ucinana, mediana lub średnia ważona. W środowisku R wyliczenie średniej dla cechy można wykonać za pomocą funkcji *mean*. Odmianą średniej jest średnia ucinana, w której odrzucana jest część danych. Eliminuje to duży wpływ na średnią wartości odstających. Zazwyczaj odrzucane jest 10% wartości najmniejszych i największych (*trim* = 0.1 w funkcji *mean*). Mediana (wartość środkowa) jest estymatorem odpornym na wartości odstające. Może być policzona z wykorzystaniem funkcji *median*. Średnia ważona może być obliczona za pomocą funkcji *weighted.mean*, której drugim argumentem jest wektor danych zawierający wagi.

*mean*(dane[['CH4']])
*mean*(dane[['CH4']]), *trim* = 0.1)
*median*(dane[['CH4']])
*weighted.mean*(dane[['CH4']]), w)

Rozproszenie (inaczej zmienność) to ocena, czy badane dane są ciasno zgrupowane, czy rozproszone. Głównymi estymatorami mogą być zakres (różnica pomiędzy wartościami największą i najmniejszą), percentyle, odchylenie międzykwartylowe i odchylenie standardowe.

Używając funkcji *summary*, można uzyskać wartości minimalne, maksymalne, średnią, medianę oraz 75. i 25. percentyl (czyli 1. i 3. kwartyl) dla analizowanych danych (dla każdej cechy z ramki danych).

Odchylenie międzykwartylowe (ang. *interquartile range*, IQR) to różnica pomiędzy 75. a 25. percentylem. Wartość IQR jest wykorzystywana do szybkiej identyfikacji wartości odstających – wartości powyżej i poniżej 1,5 · IQR są klasyfikowane jako odstające (np. na wykresach pudełkowych – boxplotach).

*summary*(dane)
*sd*(dane[['CH4']])
*IQR*(dane[['CH4']])
*mad*(dane[['CH4']])

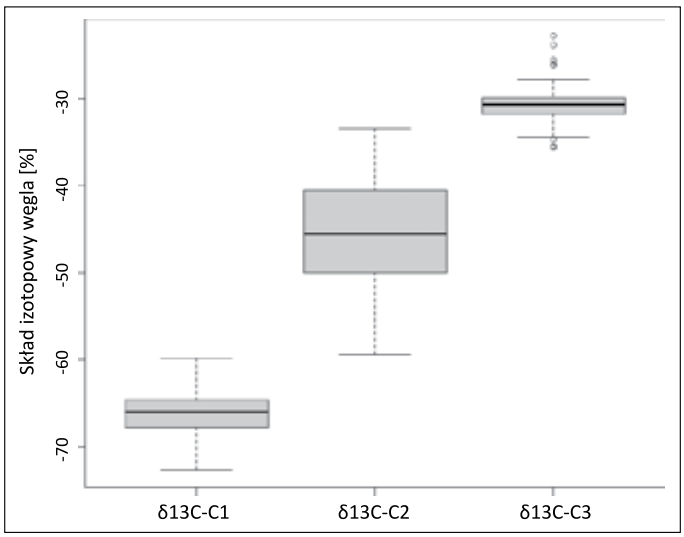
### *Wykres pudełkowy (boxplot)*

Wykres typu pudełkowego (równie często nazywany wykresem boxplot) jest bardzo użytecznym sposobem prezentacji danych. Wykres jednocześnie zawiera medianę, kwartyle oraz wartości odstające. W przypadku domyślnych ustawień w środowisku R wartości odstające to te poza zakresem 1,5 · IQR. Do wykreślania wykresów boxplot wykorzystuje się funkcję *boxplot*. Wśród wielu funkcjonalności tej funkcji można wymienić opisy wykresu oraz osi, legendę i zmianę zakresu określania wartości odstających.

Używając kodu poniżej, można wykreślić boxplot dla wartości składu izotopowego węgla w metanie, etanie i propanie (rysunek 1). Najpierw tworzony jest wektor z nazwami serii danych, następnie sam wykres.

W nawiasie podane zostały trzy serie danych do wykorzystania, opis osi y oraz nazwy serii danych (czyli wykorzystanie wektora „nazwy").

nazwy <- c('δ13C-C1', 'δ13C-C2', 'δ13C-C3')
*boxplot*(dane$`δ13C-C1`, dane$`δ13C-C2`, dane$`δ13C-C3`,
*ylab* = 'Skład izotopowy węgla [‰]', *names* = nazwy)

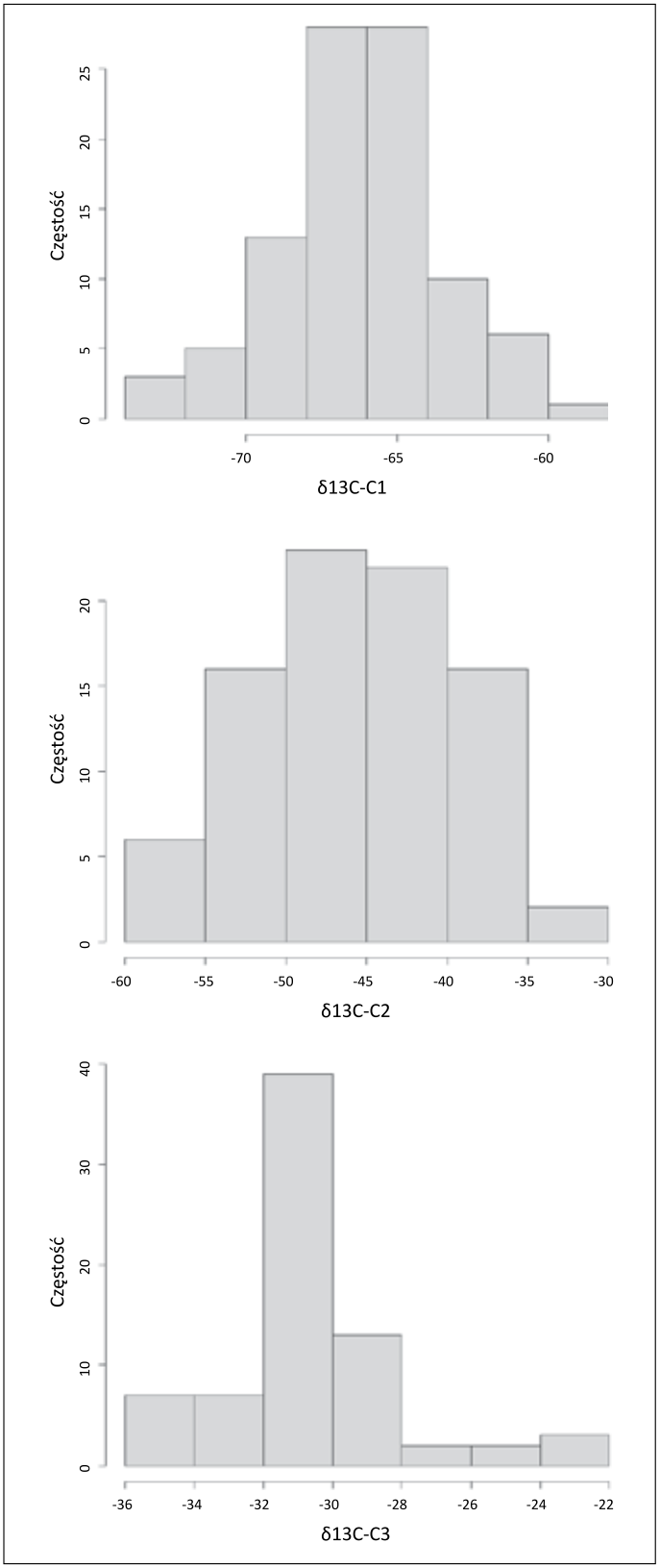



**Rysunek 1.** Wykres pudełkowy (boxplot) składu izotopowego węgla w metanie, etanie i propanie

**Figure 1.** Box plot of carbon isotopic composition in methane, ethane and propane

### *Tablica częstości i histogramy*

Rozkłady częstości można analizować, używając funkcji *cut* i *table* dla stworzenia tabeli. Bardziej czytelny sposób to stworzenie wykresów częstości, czyli histogramów. Histogramy można wykreślać przy użyciu funkcji *hist*. W kodzie poniżej tworzone są histogramy z serii danych, z opisem osi i bez tytułu wykresu (rysunek 2).




**Rysunek 2.** Histogramy (rozkłady częstości) dla wartości składu izotopowego węgla w metanie, etanie i propanie

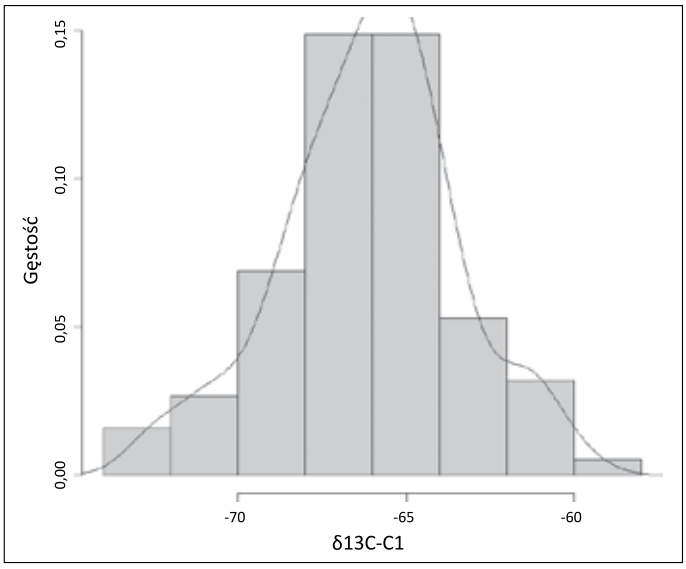
**Figure 2.** Histograms (frequency distributions) for carbon isotopic composition values of methane, ethane and propane

*hist*(dane$`δ13C-C1` , *xlab* = 'δ13C-C1' , *main* = NULL)
*hist*(dane$`δ13C-C2` , *xlab* = 'δ13C-C2' , *main* = NULL)
*hist*(dane$`δ13C-C3` , *xlab* = 'δ13C-C3' , *main* = NULL)

### *Szacowanie gęstości*

Wykres gęstości stanowi dodatek do histogramów, pozwala precyzyjniej zobrazować rozkład punktów danych. Wykres (rysunek 3) można uzyskać, łącząc funkcje *hist* oraz *lines* i *density*. Funkcja *lines* dodaje linię do już istniejącego wykresu, a jej argumentem jest funkcja *density*, która wylicza gęstość poprzez estymator jądrowy gęstości (Duong, 2001).

*hist*(dane$`δ13C-C1`, *freq* = FALSE , *xlab* = 'δ13C-C1' , *main* = NULL)
*lines*(density(dane$`δ13C-C1`), *lwd* = 1)



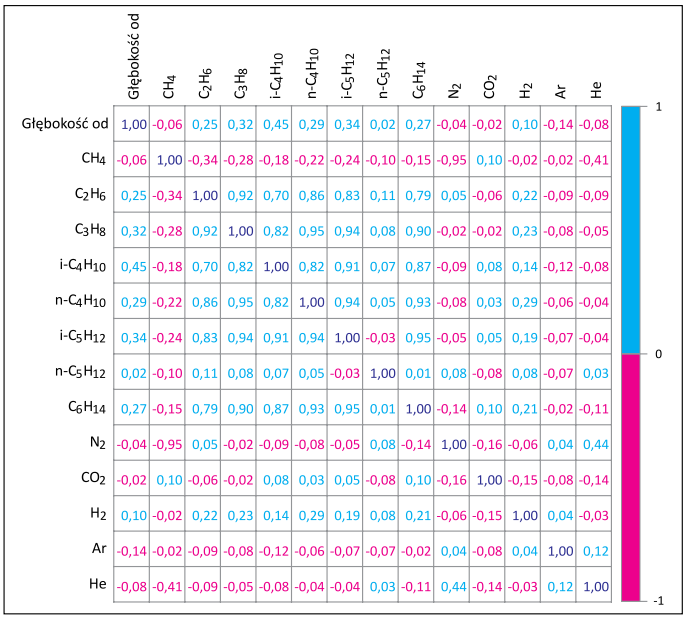

**Rysunek 3.** Histogram z wykresem gęstości rozkładu dla wartości składu izotopowego węgla w metanie

**Figure 3.** Histogram with distribution density plot for values of carbon isotopic composition in methane

### *Korelacja*

Współczynniki korelacji Pearsona są bardzo wygodne dla dostrzeżenia zależności pomiędzy różnymi zmiennymi. Zazwyczaj prezentuje się je w formie tabelarycznej. W celu stworzenia macierzy korelacji niezbędne jest zainstalowanie biblioteki *corrplot*. Podstawową funkcją wyliczającą i zwracającą współczynniki korelacji jest *cor*. Funkcja ta pozwala również wyliczyć współczynniki korelacji Kendalla i Spearmana. Funkcja *corrplot* dodatkowo umożliwia ciekawą wizualizację tabeli. Funkcja ma wiele argumentów pozwalających modyfikować wygląd na wiele sposobów (rysunek 4).

*library*(corrplot)
*cor*(danekor)
*corrplot*(cor(danekor), *method* = 'number')




| | Głębokość od | $CH_4$ | $C_2H_6$ | $C_3H_8$ | i-$C_4H_{10}$ | n-$C_4H_{10}$ | i-$C_5H_{12}$ | n-$C_5H_{12}$ | $C_6H_{14}$ | $N_2$ | $CO_2$ | $H_2$ | Ar | He |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Głębokość od | 1,00 | -0,06 | 0,25 | 0,32 | 0,45 | 0,29 | 0,34 | 0,02 | 0,27 | -0,04 | -0,02 | 0,10 | -0,14 | -0,08 |
| $CH_4$ | -0,06 | 1,00 | -0,34 | -0,28 | -0,18 | -0,22 | -0,24 | -0,10 | -0,15 | -0,95 | 0,10 | -0,02 | -0,02 | -0,41 |
| $C_2H_6$ | 0,25 | -0,34 | 1,00 | 0,92 | 0,70 | 0,86 | 0,83 | 0,11 | 0,79 | 0,05 | -0,06 | 0,22 | -0,09 | -0,09 |
| $C_3H_8$ | 0,32 | -0,28 | 0,92 | 1,00 | 0,82 | 0,95 | 0,94 | 0,08 | 0,90 | -0,02 | -0,02 | 0,23 | -0,08 | -0,05 |
| i-$C_4H_{10}$ | 0,45 | -0,18 | 0,70 | 0,82 | 1,00 | 0,82 | 0,91 | 0,07 | 0,87 | -0,09 | 0,08 | 0,14 | -0,12 | -0,08 |
| n-$C_4H_{10}$ | 0,29 | -0,22 | 0,86 | 0,95 | 0,82 | 1,00 | 0,94 | 0,05 | 0,93 | -0,08 | 0,03 | 0,29 | -0,06 | -0,04 |
| i-$C_5H_{12}$ | 0,34 | -0,24 | 0,83 | 0,94 | 0,91 | 0,94 | 1,00 | -0,03 | 0,95 | -0,05 | 0,05 | 0,19 | -0,07 | -0,04 |
| n-$C_5H_{12}$ | 0,02 | -0,10 | 0,11 | 0,08 | 0,07 | 0,05 | -0,03 | 1,00 | 0,01 | 0,08 | -0,08 | 0,08 | -0,07 | 0,03 |
| $C_6H_{14}$ | 0,27 | -0,15 | 0,79 | 0,90 | 0,87 | 0,93 | 0,95 | 0,01 | 1,00 | -0,14 | 0,10 | 0,21 | -0,02 | -0,11 |
| $N_2$ | -0,04 | -0,95 | 0,05 | -0,02 | -0,09 | -0,08 | -0,05 | 0,08 | -0,14 | 1,00 | -0,16 | -0,06 | 0,04 | 0,44 |
| $CO_2$ | -0,02 | 0,10 | -0,06 | -0,02 | 0,08 | 0,03 | 0,05 | -0,08 | 0,10 | -0,16 | 1,00 | -0,15 | -0,08 | -0,14 |
| $H_2$ | 0,10 | -0,02 | 0,22 | 0,23 | 0,14 | 0,29 | 0,19 | 0,08 | 0,21 | -0,06 | -0,15 | 1,00 | 0,04 | -0,03 |
| Ar | -0,14 | -0,02 | -0,09 | -0,08 | -0,12 | -0,06 | -0,07 | -0,07 | -0,02 | 0,04 | -0,08 | 0,04 | 1,00 | 0,12 |
| He | -0,08 | -0,41 | -0,09 | -0,05 | -0,08 | -0,04 | -0,04 | 0,03 | -0,11 | 0,44 | -0,14 | -0,03 | 0,12 | 1,00 |

**Rysunek 4.** Macierz korelacji dla wyników analiz składu chemicznego

**Figure 4.** Correlation matrix for the results of chemical composition analyses

## Analiza skupień

Analiza skupień to zbiór metod wielowymiarowej analizy statystycznej służących do wyodrębniania jednorodnych podzbiorów obiektów badanej populacji. Taka analiza jest jedną z najskuteczniejszych metod segmentacji danych. Metody te charakteryzują się możliwością porównywania i kategoryzowania obiektów opisanych za pomocą wielu atrybutów. W zależności od szczegółowych rozwiązań procedury te pozwalają na tworzenie grup (skupisk) obiektów, które są „najmniej oddalone od siebie" lub „najbardziej do siebie podobne". Obiekty te są uważane za punkty w przestrzeni wielowymiarowej, w której wymiar przestrzeni określa liczba zmiennych opisujących dane obiekty.

Ze względu na stosowane metody przetwarzania danych można wyróżnić następujące typy analiz skupień: hierarchiczne, k-średnich, k-medoidów i optymalizacyjno-iteracyjne (Kwilosz et al., 2022).

## Grupowanie metodą k-średnich (centroidów)

Grupowanie metodą k-średnich (ang. *k-means*) polega na wstępnym podzieleniu populacji na z góry założoną liczbę klas. Następnie uzyskany podział jest poprawiany poprzez przenoszenie niektórych elementów do innych klas dla osiągnięcia lepszego podziału (minimalizacja wariancji wewnątrz grup). Założona liczba klas się nie zmienia. Wykonanie grupowania

można zobrazować wykresem, nie dendrogramem (Hartigan i Wong, 1979).

Kolejność działania algorytmu, po losowym wyborze środków (centroidów) klas (skupień), to:

- przypisanie punktów do najbliższych centroidów;
- wyliczenie nowych środków skupień;
- powtarzanie algorytmu aż do osiągnięcia kryterium zbieżności (najczęściej jest to krok, w którym nie zmieniła się przynależność punktów do klas) (Gordon, 1999).

W środowisku R grupowanie metodą k-średnich możne być wykonane przy użyciu funkcji *kmeans*. Funkcja ta dzieli dane poprzez minimalizację sumy kwadratów odległości każdego z rekordów od średniej obliczonej dla danego klastra. Jako średnią klastra rozumiemy wektor średnich ze zmiennych.

Do wykonania grupowania metodą k-średnich dane muszą być kompletne. Dla każdego wiersza muszą istnieć wartości każdej cechy. Z tego względu dane z analiz gazów pochodzących z utworów miocenu autochtonicznego zapadliska przedkarpackiego zostały podzielone na dwie części: wyniki analiz składu chemicznego (94 wiersze) oraz wyniki analiz składu izotopowego wraz z wskaźnikami składu chemicznego (63 wiersze).

Kolejne kroki w programie (kod poniżej) to przypisanie nazw odwiertów do osobnego wektora (później wykorzystanego do podpisów na wykresie). Dane muszą być standaryzowane (funkcja *scale*), następnie funkcja *kmeans* przeprowadza grupowanie.

Do zobrazowania wyników można wykorzystać funkcję *fviz_cluster*, która tworzy dwuwymiarowy wykres. Osie wykresu to efekt analizy głównych składowych (PCA) i są one liniową kombinacją cech (kolumn w tabeli). Wykresy przedstawiono na rysunkach 5, 6 i 7. Na rysunku 6 dwa punkty wyraźnie odbiegają od pozostałych. Dla lepszej czytelności wartości odstające zostały usunięte oraz dodatkowo zmniejszono liczbę klastrów (rysunek 7).

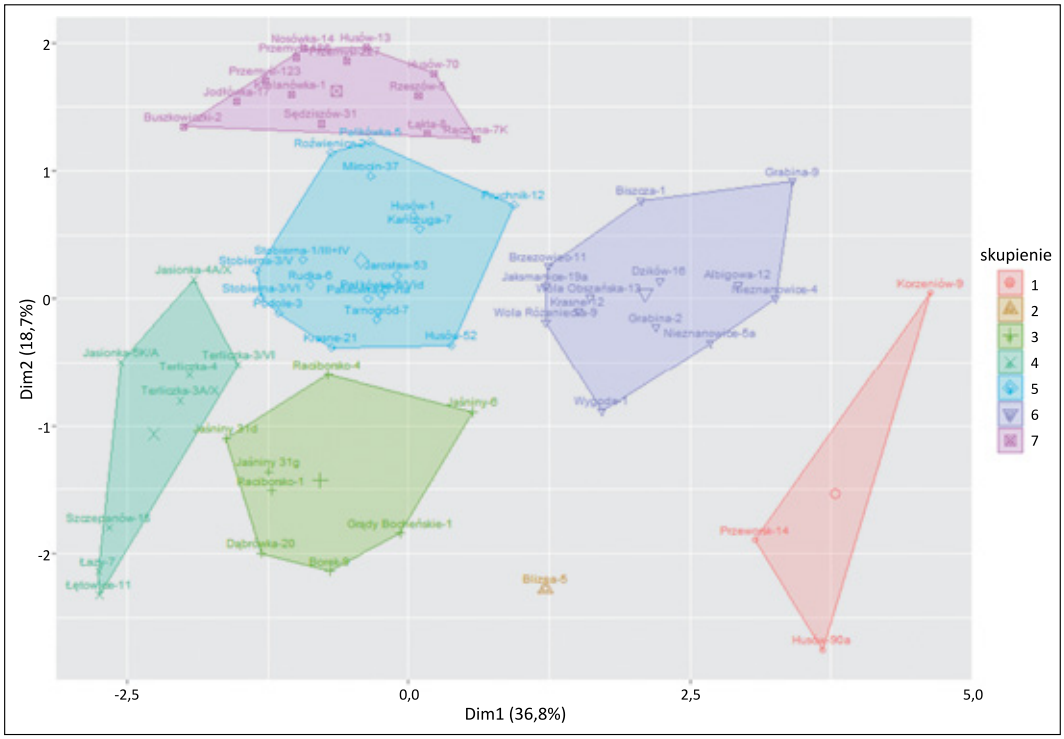
```
c <- daneizo$Odwiert
df <- daneizo
df <- df[,-1]
df <- scale(df)
rownames(df) <- c
k2 <- kmeans(df, centers = 7, nstart = 25)
fviz_cluster(k2, data = df , labelsize = 8, main = NULL)
```

## Metoda hierarchiczna

Grupowanie hierarchiczne polega na tworzeniu skupień w taki sposób, że w kolejnych iteracjach powstała grupa składa się z grup otrzymanych w poprzednim kroku. Algorytm zaczyna od pojedynczych obserwacji najbardziej podobnych do siebie i przechodzi do coraz większych grup.

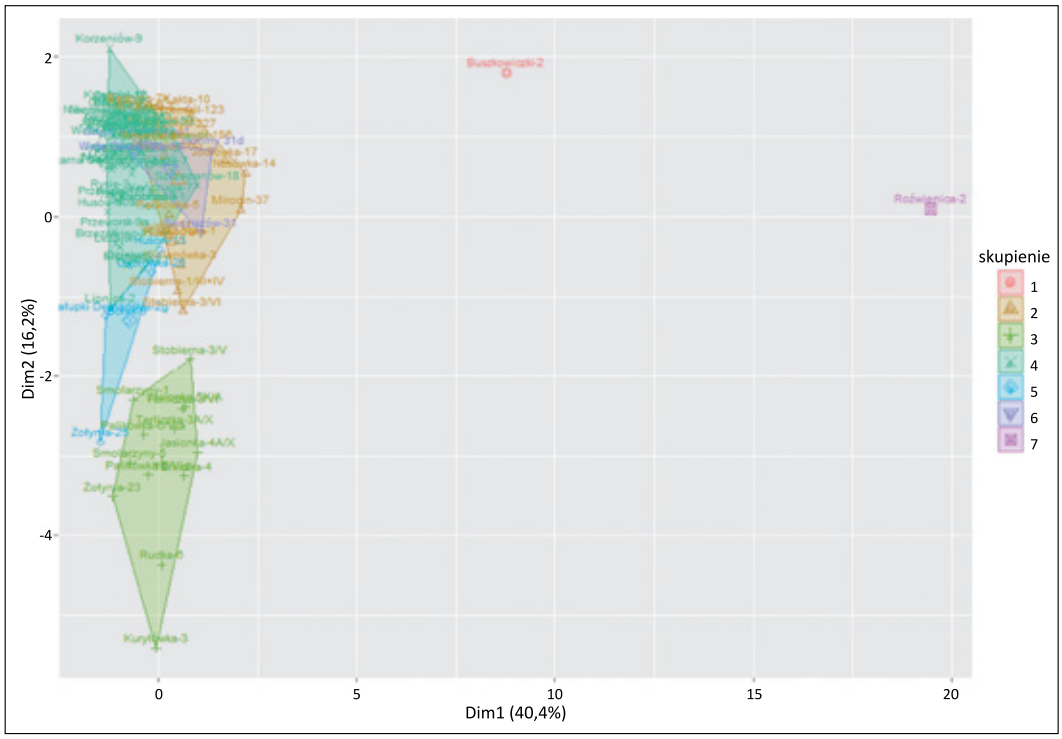
Procedury aglomeracyjne (ang. *agglomerative*) tworzą macierz podobieństwa klasyfikowanych obiektów, a następnie w kolejnych krokach łączą w skupienia obiekty (lub wcześniej utworzone grupy) najbardziej do siebie podobne. Dodatkowo brana jest pod uwagę miara podobieństwa pomiędzy klastrami (Murtagh, 1985; Murtagh i Legendre, 2014).

Grupowanie hierarchiczne prowadzi do zobrazowania wyniku w formie wykresu nazywanego dendrogramem.



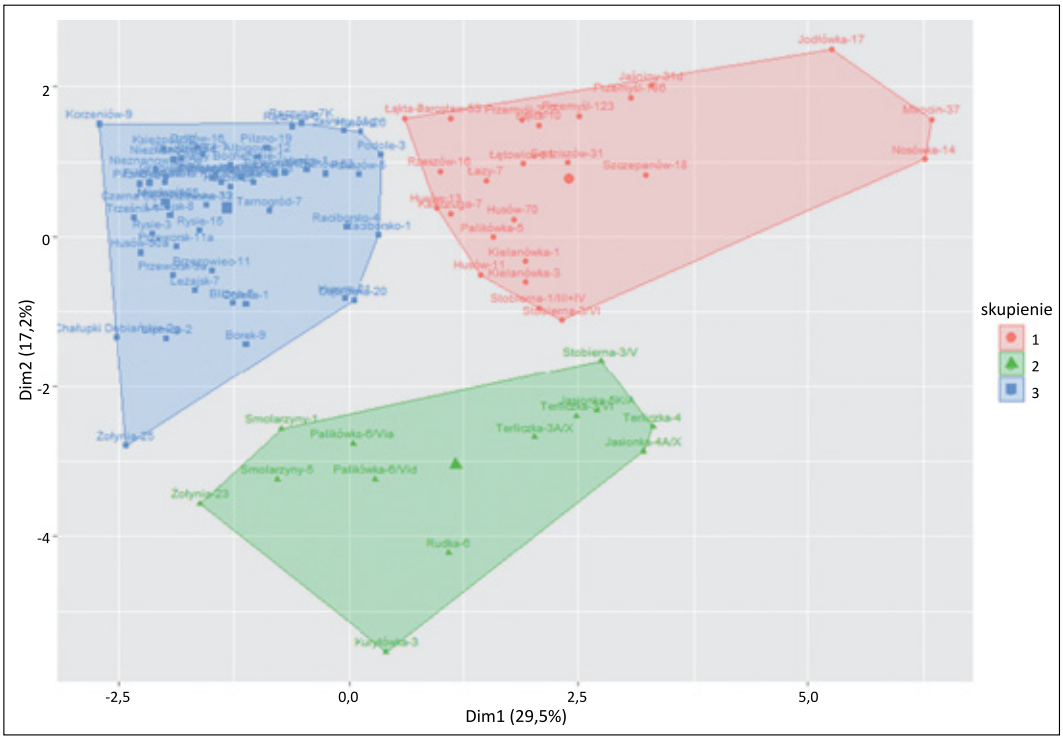

**Rysunek 5.** Wykres skupień uzyskanych metodą k-średnich – wyniki analiz składu izotopowego

**Figure 5.** Clustering diagram obtained using the k-means method – results of isotopic composition analyses




**Rysunek 6.** Wykres skupień uzyskanych metodą k-średnich – wyniki analiz składu chemicznego

**Figure 6.** Clustering diagram obtained using the k-means method – results of chemical composition analyses




**Rysunek 7.** Wykres skupień uzyskanych metodą k-średnich – wyniki analiz składu chemicznego bez wartości odstających

**Figure 7.** Cluster plot obtained with the k-means method – results of chemical composition analyses without outliers

W środowisku R do grupowania hierarchicznego można wykorzystać funkcję *hclust*. Funkcja ta wykorzystuje odległość euklidesową pomiędzy rekordami oraz miarę podobieństwa nazywaną całkowitym połączeniem.

Kolejne kroki w programie (kod poniżej) to przypisanie nazw odwiertów do osobnego wektora (później wykorzystanego do podpisów w dendrogramie). Standaryzacja (funkcja *scale*) danych poprzedza wyliczenie odległości (funkcja *dist*). Następnie funkcja *hclust* przeprowadza grupowanie hierarchiczne, a wynik jest zobrazowany dendrogramem funkcją *plot*.

```
c <- daneizo$Odwiert
df <- daneizo
df <- df[,-1]
df <- scale(df)
d <- dist(df)
hcl <- hclust(d)
```

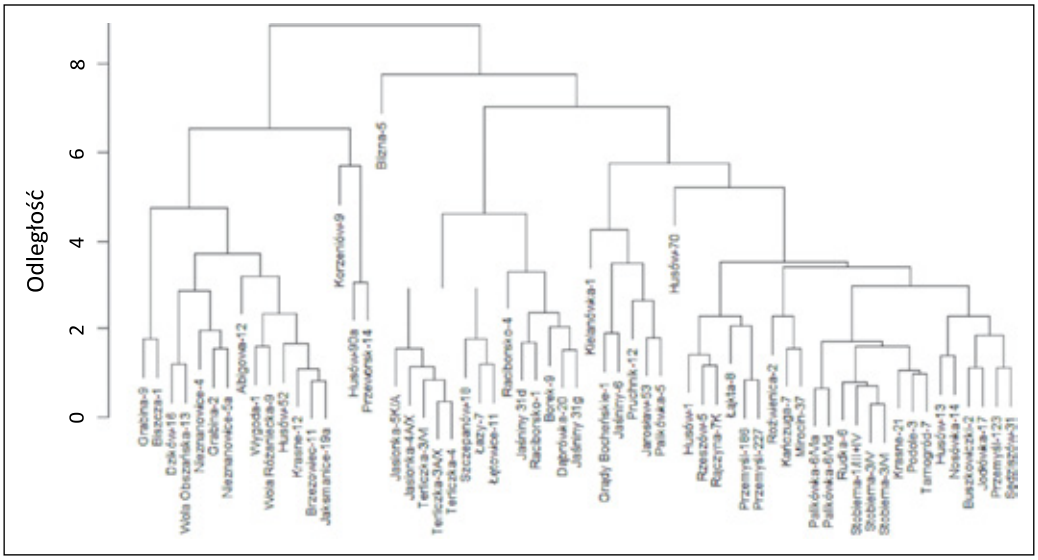
*plot*(hcl, *labels* = c, *cex* = 0.7, *ylab* = 'Odległość', *xlab* = = '', *sub* = '', *main* = '')
*cutree*(hcl, k = 8)

Dla analizowanych danych wykonano grupowanie hierarchiczne i wykreślono wykresy – dendrogramy. Dane (jak w rozdziale powyżej) były rozdzielone i osobno analizowano skład chemiczny i skład izotopowy. Wynikowe rysunki przedstawiono poniżej (rysunki 8 i 9).

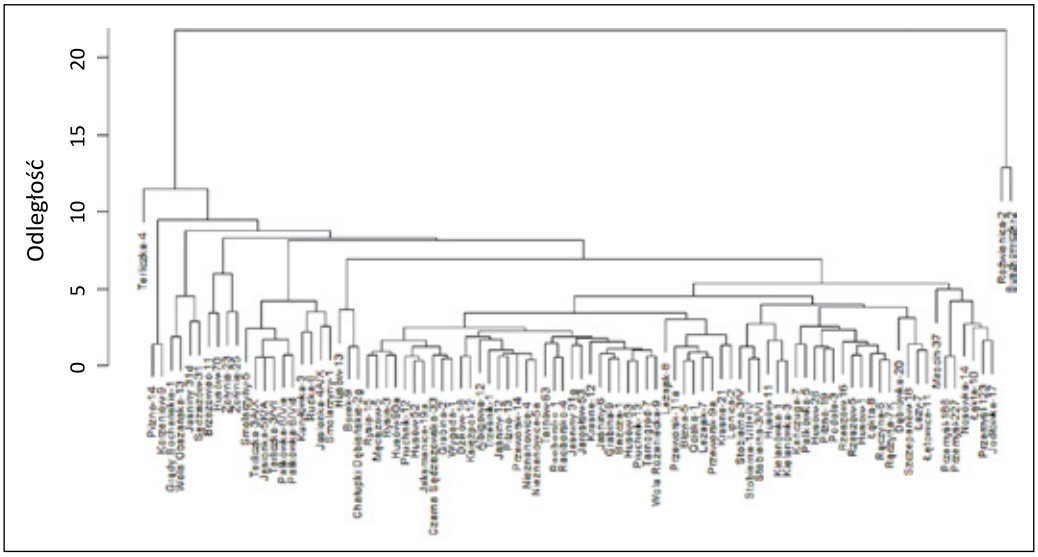
## Podsumowanie

Język R, z wykorzystaniem programu RStudio, jest bardzo wygodnym i szybkim narzędziem do statystycznej analizy danych. W przypadku geochemii naftowej, pomimo mniej licznych populacji danych, również jest możliwe wykorzystanie bardziej wyszukanych metod statystycznych. Podstawowe statystyki opisowe i wykresy (pudełkowe i histogramy) pozwalają ocenić dane i znaleźć wartości odstające. Uzyskanie wymienionych powyżej wykresów, tabel i danych jest szybkie i stosunkowo łatwe przy użyciu języka R.

Wyniki analiz składu gazu miocenu autochtonicznego wydają się mało zróżnicowane. Mimo to dzięki algorytmom k-średnich i hierarchicznym możliwe było pogrupowanie danych geochemicznych na wyraźnie rozdzielne zespoły. Zarówno wartości składu izotopowego, jak i skład chemiczny pozwalają wyznaczyć grupy, które w inny sposób nie byłyby dostrzegalne.




**Rysunek 8.** Wykres typu dendrogram – wyniki analiz składu izotopowego
**Figure 8.** Dendrogram type diagram – results of isotopic composition analyses




**Rysunek 9.** Wykres typu dendrogram – wyniki analiz składu chemicznego
**Figure 9.** Dendrogram type diagram – results of chemical composition analyses

Artykuł powstał na podstawie pracy statutowej pt. *Wykorzystanie statystycznej analizy danych, analizy skupień (HCA), do przetwarzania i interpretacji danych geochemicznych*, praca INiG – PIB, nr zlecenia: 0053/SG/2022, nr archiwalny: DK-4100-0041/2022.

## Literatura

Bruce P., Bruce A., Gedeck P., 2021. Statystyka praktyczna w data science. *Helion S.A.*

Duong T., 2001. An introduction to kernel density estimation. *Weatherburn Lecture Series, Department of Mathematics and Statistics, University of Western Australia*, 5(24).

Gordon A.D., 1999. Classification. Second Edition. *Chapman and Hall/CRC, London*.

Hartigan J.A., Wong M.A., 1979. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28: 100–108. DOI: 10.2307/2346830.

Kotarba M., 1992. Bacterial gases in Polish part of the Carpathian Foredeep and the Flysch Carpathians: isotopic and geological approach. [W:] Vially R. (ed.). Bacterial Gas. *Editions Technip, Paris*, 133–146.

Kotarba M.J., 1998. Composition and origin gaseous hydrocarbons in the Miocene strata of the Polish part of the Carpathian Foredeep. *Przegląd Geologiczny*, 46: 751–758.

Kotarba M.J., 2011. Origin of natural gases in the autochthonous Miocene strata of the Polish Carpathian Foredeep. *Annales Societatis Geologorum Poloniae*, 81: 409–424.

Kotarba M., Jawor E., 1993. Petroleum generation, migration and accumulation in the Miocene sediments and Paleozoic–Mesozoic basement complex of the Carpathian Foredeep between Cracow and Pilzno (Poland). [W:] Spencer A.M. (ed.). Generation, accumulation and production of Europe's hydrocarbons. *Special Publication of the European Association of Petroleum Geologists, 3, Springer, Heidelberg*, 295–301.

Kotarba M.J., Nagao K., 2008. Composition and origin of natural gases accumulated in the Polish and Ukrainian parts of the Carpathian region: Gaseous hydrocarbons, noble gases, carbon dioxide and nitrogen. *Chemical Geology*, 255: 426–438.

Kotarba M.J., Więcław D., Kosakowski P., Kowalski A., 2005. Hydrocarbon potential of source rocks and origin of natural gases accumulated in Miocene strata of the Carpathian Foredeep in Rzeszów area. *Przegląd Geologiczny*, 53: 67–76.

Kwilosz T., Filar B., Miziołek M., 2022. Use of Cluster Analysis to Group Organic Shale Gas Rocks by Hydrocarbon Generation Zones. *Energies*, 15(4): 1464. DOI: 10.3390/en15041464.

Murtagh F., 1985. Multidimensional Clustering Algorithms. *COMPSTAT Lectures 4. Physica-Verlag, Würzburg*.

Murtagh F., Legendre P., 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31: 274–295. DOI: 10.1007/s00357-014-9161-z.

Topór T., 2020. An integrated workflow for MICP-based rock typing: A case study of a tight-gas sandstone reservoir in the Baltic Basin (Poland). *Nafta-Gaz*, 76(4): 219–229. DOI: 10.18668/NG.2020.04.01.

Topór T., 2021. Application of machine learning algorithms to predict permeability in tight sandstone formations. *Nafta-Gaz*, 77(5): 283–292. DOI: 10.18668/NG.2021.05.01.

Tukey J., 1962. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1): 1–67.

Tukey J., 1977. Exploratory data analysis. *Addison Wesley*.

Dr inż. Marek JANIGA
Adiunkt w Zakładzie Geologii i Geochemii
Instytut Nafty i Gazu – Państwowy Instytut Badawczy
ul. Lubicz 25 A
31-503 Kraków
E-mail: *marek.janiga@inig.pl*