autorzy: **Stanisław Koter[1], Klaudia Wesołowska[2]**
[1]Uniwersytet Mikołaja Kopernika, Toruń, [2]Politechnika Śląska, Gliwice

Zastosowanie metody PCA do opisu wód naturalnych

W niniejszej pracy przedstawiono zastosowanie metody PCA do opisu wód naturalnych. Stwierdzono, że metoda ta pozwala na jednoznaczne rozróżnienie wód pochodzących z różnych ujęć w przestrzeni 2D. Umożliwia ona także na identyfikację próbek odbiegających od pozostałych tego samego pochodzenia.

1. Wprowadzenie

W obecnych czasach jakość wód przeznaczonych do celów konsumpcyjnych i na potrzeby gospodarcze powinna odpowiadać nie tylko wymogom Rozporządzenia MZ i OS, ale przede wszystkim wygórowanym oczekiwaniom odbiorców. W Polsce pobiera się wodę do picia z dwóch zasadniczych źródeł tj. z zasobów powierzchniowych i podziemnych. Wobec postępującej degradacji środowiska i zanieczyszczenia wód powierzchniowych znaczenie wód podziemnych jako źródła wody pitnej stale wzrasta. Ponieważ wody naturalne charakteryzuje szereg parametrów określających ich skład jonowy, a zarazem smak, stąd bardzo ważną rolę odgrywa stały monitoring jakości uzdatnianych wód.

Niniejsza praca ma na celu prezentację metody PCA (*principal component analysis*) w zastosowaniu do opisu monitorowanych wód naturalnych. PCA pozwala przedstawić zasób informacji zawarty w wielu zmiennych przy pomocy niewielkiej liczby czynników. Dzięki temu możliwa jest analiza danych w przestrzeni 2D lub 3D [1,2,3]. Podjęto próbę rozważenia zagadnienia, czy próbki wody pochodzące z różnych źródeł i pobierane w różnym czasie mają swoją szczególną cechę (charakterystykę) pozwalającą na identyfikację z danym źródłem.

2. Metodyka

Badaniom poddano twarde wody studzienne (pochodzące z Gliwic, Jaworzna i Będzina), charakteryzujące się zróżnicowanym składem jonowym, oraz dodatkowo wodę wodociągową. Wykonując oznaczenia 8 parametrów, tj. twardości ogólnej, węglanowej, węgla nieorganicznego, wapnia, magnezu, sodu, chlorków, siarczanów określono stężenia tych jonów w próbkach wód pobieranych w różnych okresach czasu. Stężenie magnezu, wapnia oraz oznaczenie twardości ogólnej i wykonano miareczkową metodą kompleksometryczną z EDTA, natomiast stężenie chlorków oznaczono metodą miareczkową Mohra. Również metoda miareczkowa posłużyła do oznaczenia twardości węglanowej. Zawartość siarczanów oznaczono za pomocą Merck'a SQ 118, a sód w badanych wodach mierzono za pomocą fotometru płomieniowego Flapho. Węgiel nieorganiczny oznaczono za pomocą analizatora węgla i azotu Multi N/C.

3. Opis metody PCA [1,2,3]

Niech wiersze macierzy danych **X** (*n* x *m*) odpowiadają próbkom, a kolumny - zmiennym objaśniającym te próbki. Jeśli macierz korelacji **X** tych zmiennych nie jest diagonalna, oznacza to, że są one ze sobą skorelowane, a zatem część informacji wnoszona przez każdą zmienną objaśniającą jest powtórzeniem informacji wnoszonej przez pozostałe zmienne. Należy zatem znaleźć nowe, nieskorelowane zmienne, związane liniowo ze zmiennymi pierwotnymi.

Ponieważ zmienne pierwotne mogą mieć różny charakter fizyczny, a także różnić się znacznie samymi wartościami, poddaje się je przedtem autoskalowaniu:

$$z_{ik} = (x_{ik} - \bar{x}_k)/s_k \qquad i=1,..,n;\ k=1,..,m \tag{1}$$

gdzie x_{ik}, z_{ik}, są odpowiednio elementami macierzy danych przed i po standaryzacji, $\bar{x}_k$ - średnią elementów w *k*-tej kolumnie, s_k - odchyleniem standardowym.

Problem polega zatem na wyrażeniu macierzy autoskalowanych danych **Z** w postaci iloczynu:

$$\mathbf{Z} = \mathbf{TP}' \tag{2}$$

gdzie **P**' jest transponowaną macierzą **P** (*m* x *m*), zwaną „loading matrix", a **T** jest macierzą szukanych nieskorelowanych zmiennych, zwaną „score matrix". Wobec braku korelacji iloczyn **T**'**T** winien dawać macierz diagonalną. Na tej podstawie można wykazać, że kolumny macierzy **P**, $\mathbf{p}_k$, są wektorami własnymi macierzy kowariancji cov(**Z**):

$$\mathrm{cov}(\mathbf{Z})\mathbf{p}_k = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z}\mathbf{p}_k = \lambda_k \mathbf{p}_k \tag{3}$$

gdzie λ_k jest wartością własną macierzy cov(**Z**), odpowiadającą wektorowi $\mathbf{p}_k$, i jest miarą ilości wariancji danych opisywanej przez *k*-tą kolumnę macierzy **T,** $\mathbf{t}_k$:

$$\lambda_k = \frac{1}{n-1}\mathbf{t}_k'\mathbf{t}_k \overset{\bar{t}_k=0}{=} \mathrm{var}(\mathbf{t}_k) \tag{4}$$

Ze względu na ortogonalność macierzy **P** (**P'P=I**) można zapisać:

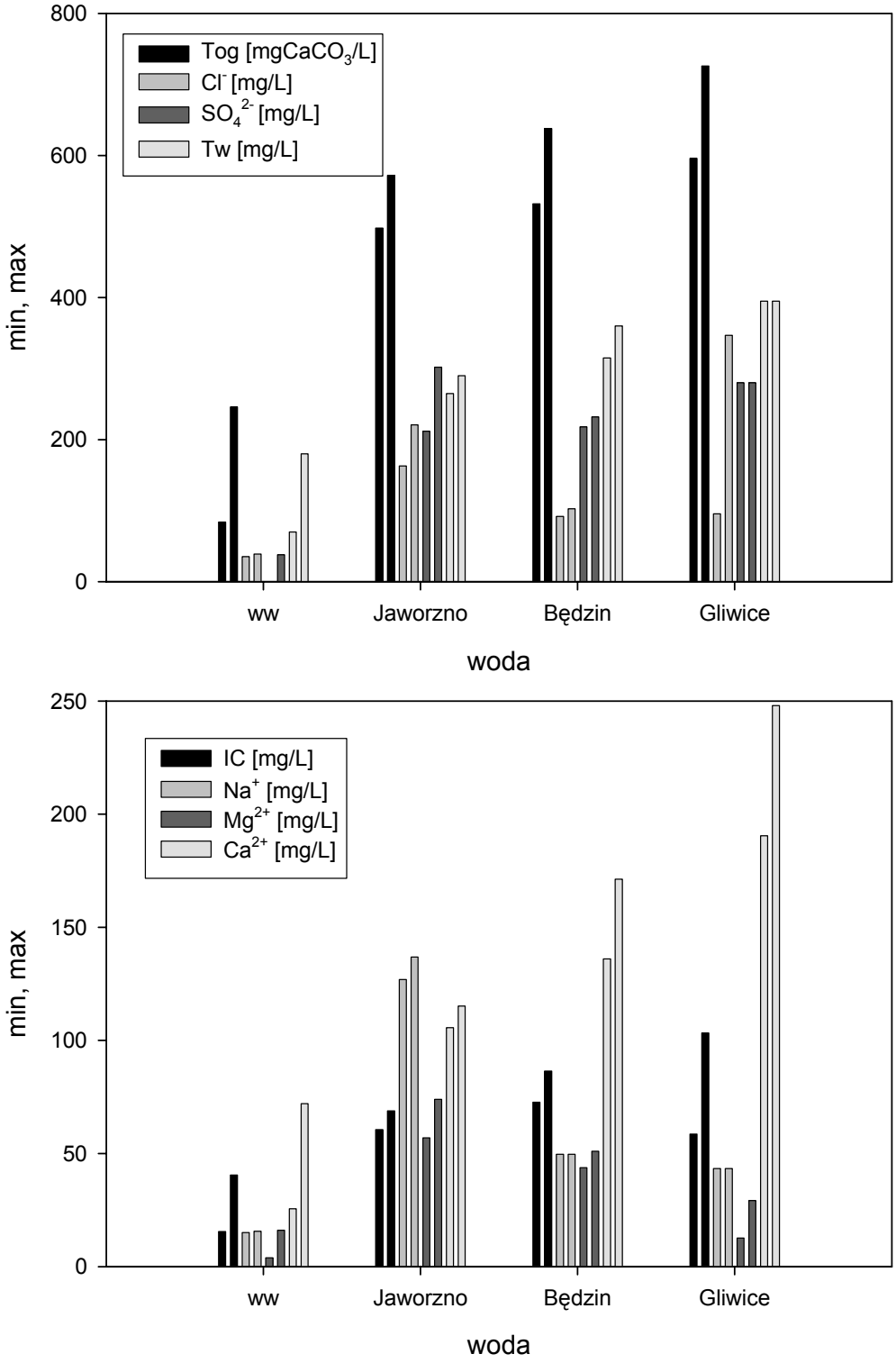
$$\mathbf{t}_k = \mathbf{Z}\mathbf{p}_k \tag{5}$$

Zgodnie z tym równaniem $\mathbf{p}_k$ określa, jaką kombinacją autoskalowanych zmiennych wyjściowych jest *k*-ta główna składowa.

Kolumny **T**, $\mathbf{t}_k$, są wzajemnie ortogonalne i zawierają współrzędne poszczególnych próbek w nowej przestrzeni tzw. głównych składowych, które przyjęto oznaczać jako PC1, PC2,... . Pierwsza główna składowa zawiera największą wariancję (zmienność) początkowego zestawu danych, każda następna składowa – coraz mniejszą. W zależności od rodzaju problemu można przyjąć, że istotne są te pierwsze składowe, które zawierają 90-95% zmienności danych, pozostałe składowe się pomija (wówczas $\mathbf{Z} = \mathbf{TP}' + \mathbf{E}$, gdzie **E** jest macierzą resztek). Dzięki temu prezentacja graficzna i analiza danych staje się znacznie łatwiejsza.

4. Dyskusja wyników i wnioski

Minimalne i maksymalne wartości parametrów uzyskane w analizowanych próbkach wód przedstawiono na rys.1.



Rys.1. Zestawienie minimalnych i maksymalnych wartości parametrów uzyskanych w przeprowadzonych analizach wód studziennych i wodociągowej

Analizując przedstawione dane można stwierdzić, iż trudno jest ustalić parametr, który pozwalałby na jednoznaczne rozróżnienie analizowanych wód. Niewątpliwie woda wodociągowa (ww) charakteryzuje się niewielkimi wartościami parametrów w porównaniu z wodami studziennymi, w obrębie których niską zawartość Mg^{2+} wykazuje woda z Gliwic (porównywalną jednak z ww). Tak więc jednoznaczne określenie pochodzenia próbki wody wcale nie jest takie proste.

Zestawienie współczynników korelacji pomiędzy poszczególnymi parametrami (zmiennymi objaśniającymi) prezentuje tab.1.

Tab.1. Macierz korelacji parametrów wody (zmiennych objaśniających), równa macierzy kowariancji parametrów po autoskalowaniu $\mathrm{cov}(\mathbf{Z}) = \mathbf{Z}'\mathbf{Z}/(n-1)$

	T_{og} [$CaCO_3$]	Cl-	SO_4^{--}	T_w	IC	Na^+	Mg^{++}	Ca^{++}
T_{og} [$CaCO_3$]	1							
Cl^-	0.569	1						
SO_4^{--}	0.946	0.703	1					
T_w	0.956	0.648	0.886	1				
IC	0.850	0.746	0.824	0.937	1			
Na^+	0.441	0.309	0.594	0.227	0.223	1		
Mg^{++}	0.614	0.196	0.656	0.414	0.403	0.869	1	
Ca^{++}	0.897	0.596	0.806	0.955	0.829	0.061	0.201	1

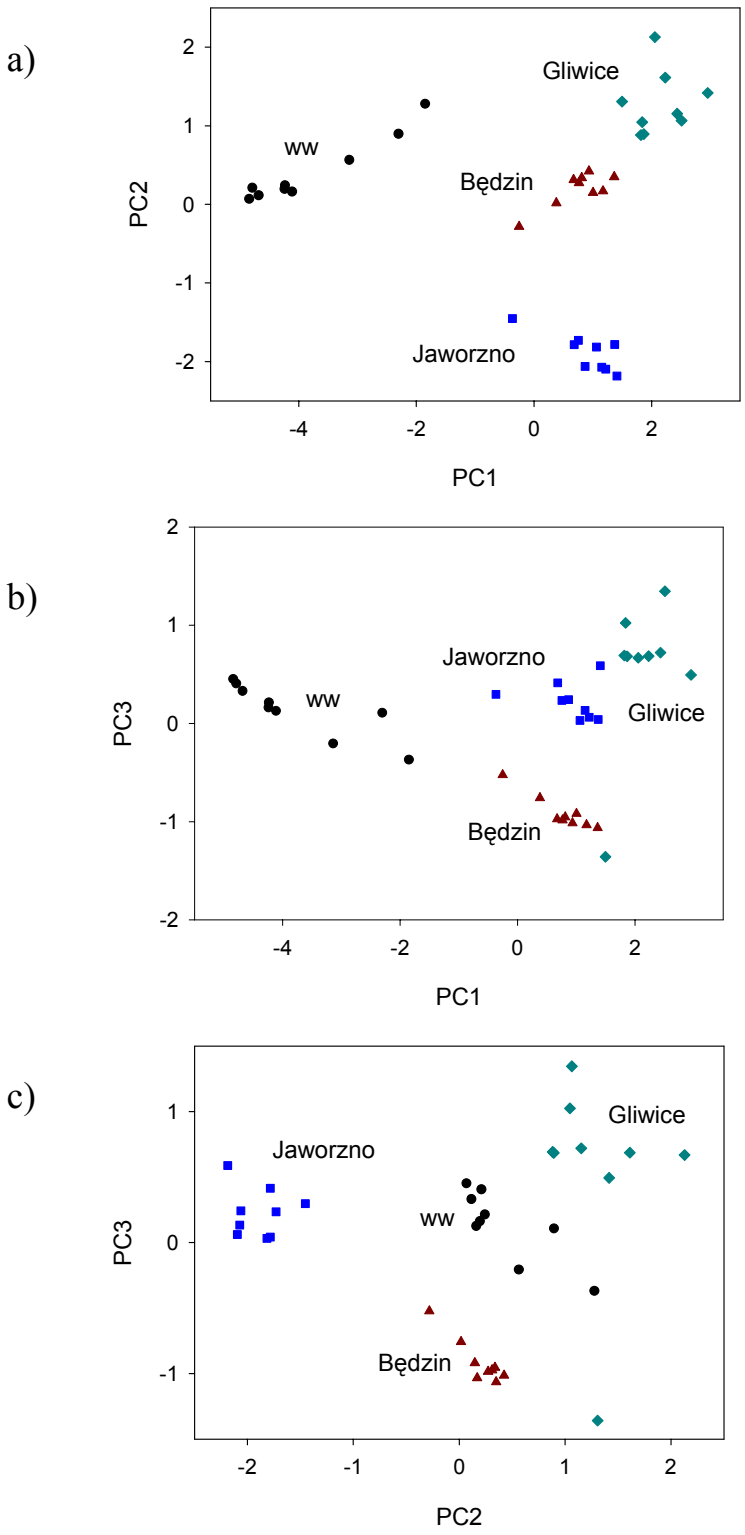
Jest widoczne, że pomiędzy tymi parametrami występuje silna korelacja (np. między T_{og} a SO_4^{2-} czy T_w), a zatem jak najbardziej uzasadnione jest przeprowadzenie analizy głównych składowych.

W tab.2 zestawiono wariancje głównych składowych, λ_k, oraz ich procentowy udział w zmienności zestawu danych.

Tab.2. Wariancja głównych składowych i ich %-wy udział w zmienności zestawu danych

PC ($\mathbf{t}_k$) k	Wariancja λ_k	% udziału
1	5.528	69.1
2	1.627	20.3
3	0.592	7.4
4	0.195	2.4
5	0.032	0.40
6	0.023	0.29
7	0.0025	0.03
8	$8.2 \cdot 10^{-8}$	0.00

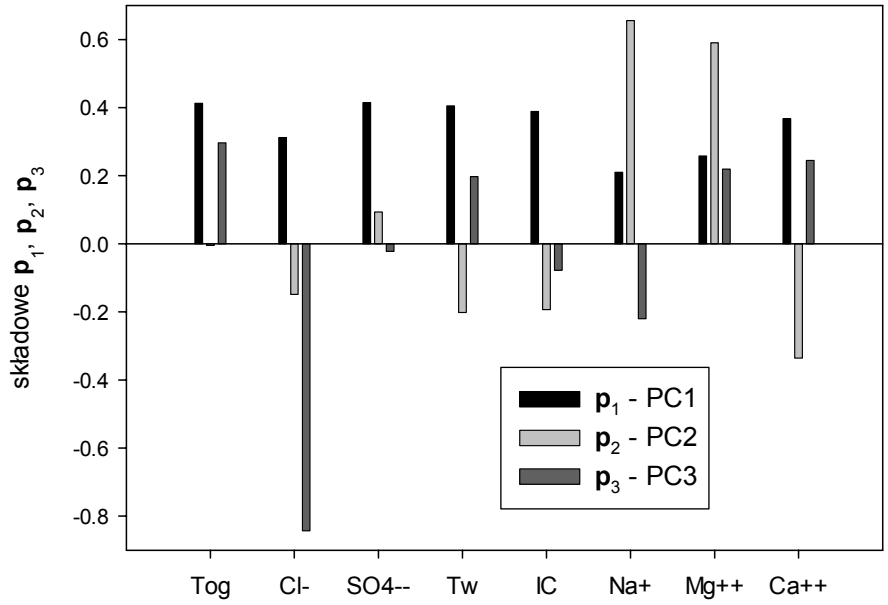
Jest widoczne, że pierwsze trzy składowe objaśniają prawie 97% zmienności zawartej w danych, przy czym pierwsza składowa objaśnia ponad 2/3 zmienności. A zatem do opisu badanych próbek wody wystarczą 2-3 pierwsze główne składowe.



Rys.2. Rzut próbek na płaszczyznę głównych składowych: a) PC1-PC2, b) PC1-PC3, c) PC2-PC3.

Na rys.2 przedstawiono próbki wody wykreślone na płaszczyźnie dwóch spośród trzech pierwszych składowych (a) PC1-PC2, b) PC1-PC3 i c) PC2-PC3). Jest widoczne, że rzut próbek na płaszczyznę PC1-PC2 pozwala na wyraźne rozróżnienie próbek, jeśli chodzi o źródło ich pochodzenia. W przypadku PC1-PC3, jak również PC2-PC3, można dostrzec wyraźnie odbiegającą próbkę w przypadku wody z Gliwic. Przyczyną tej rozbieżności może być błąd w analizie, a jeśli to nie wchodzi w rachubę, to chwilowa zmiana w składzie wody wynikająca niekoniecznie z przyczyn naturalnych.

Udział poszczególnych parametrów w głównych składowych przedstawiają wektory $\mathbf{p}_k$ (r.(5), rys.3). Zgodnie z danymi, przedstawionymi na rys.3, przybliżony sens fizyczny głównych składowych jest następujący: PC1 charakteryzuje całkowitą zawartość wszystkich składników wody, natomiast PC2 jest kombinacją przede wszystkim stężeń kationów.



Rys.3. Składowe wektorów $\mathbf{p}_1$, $\mathbf{p}_2$ i $\mathbf{p}_3$ transformujących autoskalowane parametry wody w główne składowe PC1, PC2, PC3 (r.(5))

Podsumowując, zastosowana metoda PCA spełnia swoje zadanie przyporządkowania otrzymanych wyników do odpowiednich źródeł poboru wód. Pozwala ona także na identyfikację próbek odbiegających od pozostałych tego samego pochodzenia. Rozbieżności te mogą; być spowodowane okresowymi wahaniami składu ujmowanych wód, a nawet wskazywać na „chwilowe" ich zanieczyszczenia.

5. Literatura

1. Mazerski J., „Podstawy chemometrii", Wyd. Politechniki Gdańskiej, Gdańsk, 2000.
2. Beebe K.R., Pell R.J., Seasholtz M.B., „Chemometrics. A practical guide", J. Wiley & Sons, Inc., New York, 1998.
3. Wise B.M., Gallagher N.B., J. Proc. Cont. 6(1996)329.