ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2005
Seria: GÓRNICTWO z. 267 Nr kol. 1690

Danuta WIECHUŁA*, Krzysztof LOSKA**, Jerzy KWAPULIŃSKI*
*Katedra i Zakład Toksykologii, Śląska Akademia Medyczna, Sosnowiec
**Instytut Inżynierii Wody i Ścieków, Politechnika Śląska, Gliwice

ZASTOSOWANIE PROCEDUR STATYSTYCZNYCH W INTERPRETACJI WYNIKÓW BADAŃ ŚRODOWISKOWYCH

Streszczenie. W badaniach środowiskowych statystyczne opracowanie w każdym przypadku powinna poprzedzić dokładna analiza zbioru uzyskanych wartości, zwłaszcza w przypadku, gdy w zestawie wyników znajdują się pojedyncze wartości znacznie odbiegające od większości otrzymanych wyników. Występowanie pojedynczych wartości odbiegających może być spowodowane zarówno punktowym zanieczyszczeniem, jak również występowaniem błędów przypadkowych, związanych z zanieczyszczeniem próbki na etapie pobierania, preparatyki lub oznaczania. W decyzji o odrzuceniu lub włączeniu wartości odbiegających z posiadanego zbioru danych mogą pomóc testy statystyczne zaproponowane przez EPA. W niniejszej pracy wybrane testy zastosowano do analizy wartości odbiegających w zbiorach danych zawartości miedzi w osadzie dennym zbiornika Dziećkowice. Na podstawie przeprowadzonych badań wykazano, że wartości znacznie wyższe od większości ze zbioru danych są wartościami odbiegającymi i mogą zostać odrzucone z posiadanego zbioru po wnikliwym przeanalizowaniu kolejnych etapów analizy chemicznej w poszukiwaniu źródeł przypadkowego błędu prowadzącego do uzyskania zawyżonych wartości.

APPLICATION OF STATISTICAL PROCEDURES TO THE INTERPRETATION OF ENVIRONMENTAL RESEARCH RESULTS

Summary. In environmental research, each statistical analysis should be preceded by a thorough investigation of a given data set, especially if it contains outliers. The occurrence of single outliers may result from both point source contamination and accidental errors caused by sample contamination during collection, preparation or determination steps. The decision whether to include or exclude outliers may be made with the help of statistical tests proposed by EPA. In this work, selected statistical tests were used to analyze the outliers from the data set on copper content in the bottom sediment of Dziećkowice reservoir. It has been found that the values much higher than the ones representative of the data set are outliers and can be eliminated after the applied chemical analysis has been thoroughly checked for possible accidental errors which could produce the skewed values.

1. Wprowadzenie

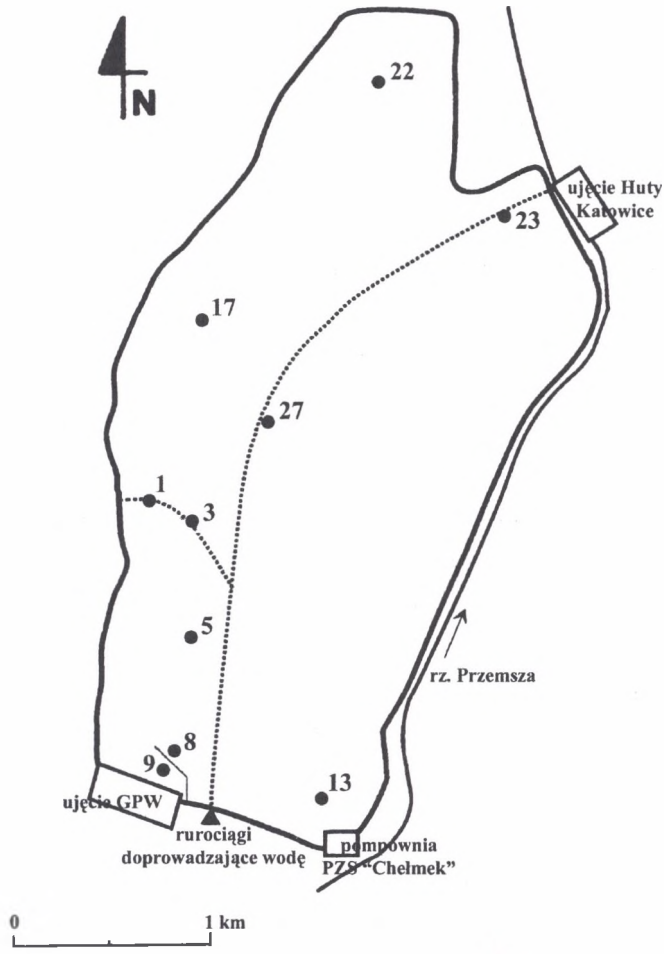
Końcowym etapem wszystkich badań środowiskowych jest interpretacja i dyskusja wyników na podstawie ich statystycznego opracowania. Wykorzystanie nawet najprostszych analiz statystycznych wymaga informacji o charakterze rozkładu uzyskanych wyników. Większość klasycznych testów statystycznych, analiza korelacyjna, analiza czynnikowa i inne są oparte na założeniu, że rozkład wyników jest normalny, dlatego też otrzymany zbiór danych powinien w każdym przypadku być uważnie analizowany. W każdym przypadku konieczne jest wyznaczenie charakterystyki rozkładu zmiennych [1, 2]. Przy analizie danych często pojawia się problem tzw. wartości odbiegających, czyli wyników znacznie różniących się od pozostałych. W każdym takim przypadku należy ustalić, czy badana zmienna ma taki charakter rozkładu, czy też występowanie wartości odbiegających jest wynikiem błędu związanego z zanieczyszczeniem próbki na etapie pobierania, preparatyki lub oznaczania. Powstaje pytanie, w jaki sposób sprawdzić, czy określony zbiór danych jest zbiorem jednorodnym oraz czy wartości odbiegające są wynikiem błędu i powinny być odrzucone ze zbioru danych?

Proste testy statystyczne pomagające odpowiedzieć na to pytanie zostały zaproponowane przez Amerykańską Agencję Ochrony Środowiska (EPA) [1, 3]. Od razu jednak pojawia się zastrzeżenie, że wynik testu nie może być jedynym powodem odrzucenia badanej wartości odbiegającej z posiadanego zbioru danych, lecz powinien prowadzić do powtórnego przeanalizowania etapów analizy chemicznej w celu znalezienia ewentualnych błędów. Test statystyczny dla wartości odbiegającej może również dawać błędną informację, ponieważ w większości przypadków oparty jest na założeniu o normalności rozkładu, podczas gdy rzeczywisty rozkład otrzymanych wyników może być np. log-normalny.

W niniejszej pracy przedstawiono praktyczne zastosowanie wybranych testów dla wartości odbiegającej, na przykładzie zbioru danych zawartości miedzi w osadzie dennym ze zbiornika Dziećkowice.

2. Materiał i metody

Osady denne, pobrane z 10 stanowisk z obszaru zbiornika Dziećkowice (rys. 1), suszono do stałej masy w temperaturze 105°C.



Rys. 1. Lokalizacja stanowisk poboru próbek na obszarze zbiornika Dziećkowice
Fig. 1. Localization of sampling points within Dziećkowice reservoir

Suchy osad denny rozdrabniano i mielono do średnicy mniejszej niż 0,01 mm. 250 mg ± 20% osadu dennego mineralizowano metodą mikrofalową mieszaniną kwasu azotowego i fluorowodorowego w ilościach odpowiednio 3 i 2 cm^3. Oprócz tego dodawano 1 cm^3 H_2O_2. Po mineralizacji i odparowaniu kwasów próbki przenoszono do kolbek miarowych o objętości 50 cm^3 i uzupełniano do kreski.

Zawartość miedzi w osadach dennych oznaczano metodą płomieniową absorpcyjnej spektrofotometrii atomowej z użyciem spektrofotometru AAS-30 produkcji Carl Zeiss Jena.

Równolegle w identyczny sposób przeprowadzano mineralizację i oznaczenie miedzi w standardowym materiale referencyjnym CRM 277 (Trace elements in estuarine sediment), uzyskując wartość 98,5 ± 0,8 mg Cu/kg (wartość certyfikowana 101,7 ± 1,6 mg Cu/kg).

Ogólny zbiór danych zawierał 334 wyniki oznaczeń zawartości miedzi w osadzie dennym. Z posiadanego zbioru wyodrębniono dwa podzbiory, jeden obejmujący stanowisko 9 (n = 19) oraz drugi - stanowisko 27 (n = 43).

3. Testy wartości odbiegających

Test Dixona, n ≤ 25 [1, 3]

Jako $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ określamy n-te oznaczenie zawartości pierwiastka w zbiorze danych, uporządkowanych od wartości najmniejszej do największej.

$x_{(n)}$ – oznacza największą wartość, znacznie odbiegającą od pozostałych (wartość odbiegająca).

W pierwszej kolejności przeprowadzamy test normalności dla zbioru danych z wyłączeniem wartości $x_{(n)}$ i ustalamy poziom istotności α. W teście Dixona α może wynosić 0,01, 0,05 lub 0,10.

Obliczamy:

$C = [x_{(n)} - x_{(n-1)}] / [x_{(n)} - x_{(1)}]$, jeżeli $3 \leq n \leq 7$

$C = [x_{(n)} - x_{(n-1)}] / [x_{(n)} - x_{(2)}]$, jeżeli $8 \leq n \leq 10$

$C = [x_{(n)} - x_{(n-2)}] / [x_{(n)} - x_{(2)}]$, jeżeli $11 \leq n \leq 13$

$C = [x_{(n)} - x_{(n-2)}] / [x_{(n)} - x_{(3)}]$, jeżeli $14 \leq n \leq 25$

jeżeli C przewyższa krytyczną wartość tabelaryczną [1] dla wyszczególnianego n i α, wówczas stwierdzamy, że $x_{(n)}$ jest wartością odbiegającą.

Test Rosnera, n ≥ 25 [1, 3]

Krok 1

Wyznaczamy poziom istotności α.

$x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ wskazuje n-te oznaczenie zawartości pierwiastka w zbiorze danych, uszeregowanych w porządku od najmniejszego do największego, gdzie $n \geq 25$.

Następnie wskazujemy maksymalną liczbę wartości odbiegających – r.

Krok 2

Przyjmujemy, że $i = 0$ i obliczamy następujące równania:

$$\bar{x}^{(i)} = (x_i + x_2 + ... + x_{n-1})/(n-1)$$

$$s^{(i)} = \{[(x_i - \bar{x}^{(i)})^2 + (x_2 - \bar{x}^{(i)})^2 + ... + (x_{n-1} - \bar{x}^{(i)})^2]/(n-i)\}^{1/2}$$

Obliczamy średnią arytmetyczną, oznaczoną jako $\bar{x}^{(0)}$ i $s^{(0)}$ z wszystkich n pomiarów. Wybieramy pomiar, który jest najdalszy od $\bar{x}^{(0)}$ i oznaczamy go jako $y^{(0)}$.

Eliminujemy $y^{(0)}$ ze zbioru danych i obliczamy (używając $i = 1$ w powyższym równaniu) średnią arytmetyczną podzbioru, oznaczoną jako $\bar{x}^{(1)}$ i $s^{(1)}$ z pozostałych $n-1$ pomiarów. Wybieramy pomiar, który jest najdalszy od $\bar{x}^{(1)}$ i oznaczamy go jako $y^{(1)}$.

Eliminujemy $y^{(1)}$ ze zbioru danych i obliczamy (używając $i = 2$ w powyższym równaniu) średnią arytmetyczną, oznaczaną jako $\bar{x}^{(2)}$ i $s^{(2)}$ z pozostałych $n-2$ pomiarów.

Kontynuujemy ten proces aż do momentu, kiedy zostanie ze zbioru danych usunięte r największych pomiarów.

Krok 3

Dla sprawdzenia, czy dana wartość jest wartością odbiegającą, obliczamy:

$$R_r = [|\, y^{(r-1)} - \bar{x}^{(r-1)} \,|]/s^{(r-1)}$$

Określamy tabelaryczną wartość krytyczną λ_r [1] dla wartości n, r, i α.

Jeżeli $R_r > \lambda_r$, wartość r jest wartością odbiegającą w zbiorze danych.

Jeżeli nie, przeprowadzamy test dla kolejnej wartości odbiegającej $r-1$, obliczając:

$$R_{r-1} = [|\, y^{(r-2)} - \bar{x}^{(r-2)} \,|]/s^{(r-2)}$$

Określamy tabelaryczną wartość krytyczną λ_{r-1} dla wartości n, $r-1$ i α.

Jeżeli $R_{r-1} > \lambda_{r-1}$, wartość $r-1$ jest wartością odbiegającą w zbiorze danych.

Postępowanie powtarzamy aż do określenia, że w zbiorze danych jest pewna liczba wartości odbiegających lub że żadne wartości odbiegające nie występują w zbiorze.

Test Walsha, n > 60 [1, 3]

Niech $x_{(1)}, x_{(2)}, ..., x_{(n)}$ wskazuje n-te oznaczenie zawartości pierwiastka w zbiorze danych, uszeregowanych od najmniejszego do największego. Jeżeli $60 < n \leq 220$, przyjmujemy $\alpha = 0{,}10$, jeżeli $n > 220$ - $\alpha = 0{,}05$.

Określamy liczbę prawdopodobnych wartości odbiegających, r.

Obliczamy:

$c = [(2n)^{1/2}]$.

W powyższym równaniu nawias kwadratowy oznacza wartość zaokrągloną do największej możliwej całości.

$k = r + c$,

$b^2 = 1/\alpha$

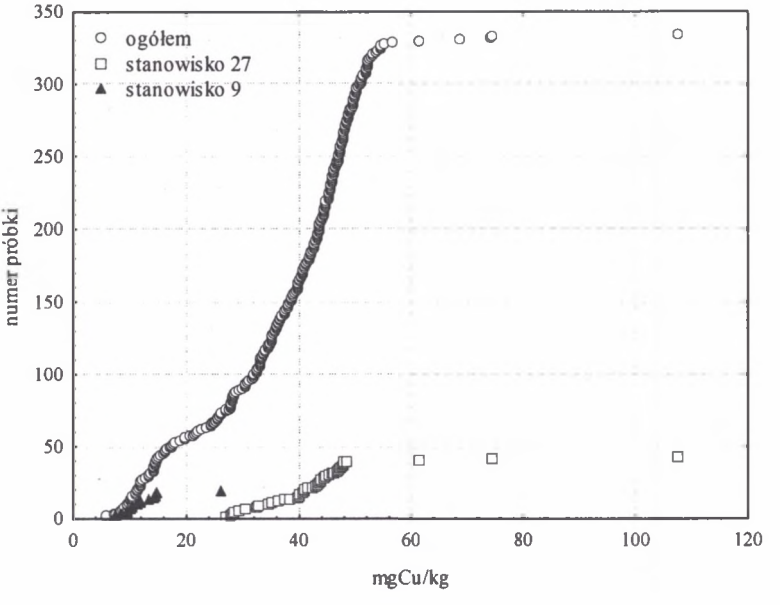
$a = (1 + b\{(c-b^2) / (c-1)\}^{1/2}) / (c - b^2 - 1)$

Największe pomiary r są wartościami odbiegającymi (na poziomie istotności α), jeżeli:

$x_{(n+1-r)} - (1 + a)x_{(n-r)} + ax_{(n+1-k)} > 0$

4. Wyniki

Rozkład uzyskanych wyników zawartości miedzi w osadzie dennym zbiornika Dziećkowice przedstawiono na rys. 2.



Rys. 2. Zawartość miedzi w osadzie dennym zbiornika Dziećkowice
Fig. 2. Copper contetnt in sediments of Dziećkowice reservoir

Analiza danych dla całego zbiornika wskazuje, że zawartość 107,40 mg Cu/kg jest wartością znacznie większą od pozostałych i może być wartością odbiegającą. Posiadany

zbiór danych zawiera 334 wartości; w celu stwierdzenia, czy dana wartość należy do zbioru, przeprowadzono test Walsha.

W tym celu obliczono:

$c = [(2*334)^{1/2}] = 26$

$k = 1 + 26 = 27,$

$b^2 = 1/0,05 = 20$

$a = (1 + 4,47 \{(26-20) / (26-1)\}^{1/2}) / (26 - 20 - 1) = 0,64$

$x_{(n + 1 - r)} = x_{(334+1-1)} = x_{(334)} = 107,40$

$x_{(n - r)} = x_{(334-1)} = x_{(333)} = 74,42$

$x_{(n + 1 - k)} = x_{(334 + 1 - 27)} = x_{(n + 1 - k)} = x_{(308)} = 51,80$

i ostatecznie:

$x_{(n + 1 - r)} - (1 + a)x_{(n - r)} + ax_{(n + 1 - k)} = 107,40 - (1 + 0,64)*74,42 + 0,64*51,80 = 18,54$

Ponieważ obliczona wartość jest większa od 0, zgodnie z interpretacją testu Walsha można przyjąć, że wartość 107,40 mg/kg jest wartością odbiegającą.

Zbiór wyników zawartości miedzi w osadzie dennym na stanowisku 27 zawierał 43 wartości, z których trzy były wyraźnie większe od pozostałych – 61,58 mg/kg, 74,42 mg/kg oraz 107,40 kg/kg (rys. 2). Do oceny, czy dane wartości są wartościami odbiegającymi, zastosowano test Rosnera.

W tym celu obliczono średnią i odchylenie standardowe dla danych wyjściowych oraz po eliminacji kolejnych wartości odbiegających:

I	$\bar{x}^{(i)}$	$s^{(i)}$	$y^{(i)}$
0	42,75	13,34	107,40
1	41,21	8,97	74,42
2	40,40	7,40	61,58

Następnie sprawdzono, czy dane trzy największe wartości są wartościami odbiegającymi, obliczając R_3 z równania:

$R_3 = |y^{(2)} - \bar{x}^{(2)}|/s^{(2)} = |61,58 - 40,40|/7,40 = 2,86$

Wartość odczytana z tabeli dla n = 40, r = 3 i $\alpha = 0,05$ wynosi $\lambda_r = 3,01$. $R_3 = 2,86$ jest mniejsza od wartości tabelarycznej, stąd badane 3 wartości nie są wartościami odbiegającymi. W dalszej kolejności sprawdzono, czy dane 2 największe wartości są wartościami odbiegającymi, obliczając R_2 z równania: $R_2 = |y^{(1)} - \bar{x}^{(1)}|/s^{(1)} = |74,42 - 41,21|/8,97 = 3,70$.

Wartość odczytana z tabeli dla n = 40, r = 2 i $\alpha = 0,05$ wynosi $\lambda_r = 3,03$. $R_2 = 3,70$ jest większa od wartości tabelarycznej, a zatem badane 2 wartości są wartościami odbiegającymi.

Ostateczny wynik testu wskazuje, że zawartości 74,42 mg/kg i 107,40 mg/kg można uważać za wartości odbiegające, natomiast nie ma podstaw, aby odrzucić zawartość 61,58 mg/kg ze zbioru danych.

Z 19 wyników oznaczenia zawartości miedzi na stanowisku 9 jedna jest wyraźnie większa od pozostałych (rys. 2). Czy można ją traktować jako odbiegającą, sprawdzono za pomocą testu Dixona.

Dla n = 19 obliczono:

$C = [x_{(n)} - x_{(n-2)}] / [x_{(n)} - x_{(3)}] = [x_{(19)} - x_{(17)}] / [x_{(19)} - x_{(3)}] = (25,97 - 14,57)/(25,97 - 8,45) = 11,40/17,52 = 0,65$

Odczytana tabelaryczna wartość dla n = 19 i $\alpha = 0,05$ wynosi 0,462.

Ponieważ C = 0,65 przewyższa krytyczną wartość tabelaryczną 0,462, możemy stwierdzić, że $x_{(n)} = 25,97$ mg/kg w posiadanym zbiorze danych jest wartością odbiegającą.

Wyniki testów wskazują, że w zbiorze zawartości miedzi w osadzie dennym występują wartości odbiegające. Zgodnie z zaleceniami EPA, ich odrzucenie z posiadanego zbioru danych przed statystycznym opracowaniem wyników powinna poprzedzić wnikliwa analiza postępowania przygotowawczego i zastosowanej metody oznaczenia, które mogły być źródłem przypadkowego błędu prowadzącego do uzyskania zawyżonych wartości.

W dalszej kolejności zasadne było przeanalizowanie, w jaki sposób odrzucenie tych wartości ze zbioru danych wpływa na zmianę podstawowych parametrów statystycznych opisujących zbiory danych (tabl. 1).

Zarówno wartość średniej, jak i mediany w zbiorach danych uzyskanych po odrzuceniu wartości odbiegających były mniejsze. Najlepsze dopasowanie do rozkładu normalnego po eliminacji uzyskano dla zawartości miedzi w osadzie dennym na stanowisku 27. W tym przypadku wartości skośności i kurtozy zbliżyły się do zera, zmniejszyła się również wartość odchylenia standardowego. Także zawartość miedzi w osadzie dennym na stanowisku 9 po odrzuceniu wartości odbiegającej miała charakter normalny (test Kołmogorowa-Smirnowa, $p < 0,05$), a wartość odchylenia standardowego zmniejszyła się o połowę. Najmniejsze różnice stwierdzono po odrzuceniu wartości odbiegającej ze zbioru zawartości miedzi w osadzie dennym całego zbiornika, choć również w tym przypadku wartość kurtozy zbliżyła się do zera w porównaniu ze zbiorem danych wyjściowych.

Tablica 1

Analiza statystyczna zawartości miedzi w osadzie dennym zbiornika „Dziećkowice"

	Ogółem		Stanowisko 9		Stanowisko 27	
	A, n=334	B, n=333	A, n=19	B, n=18	A, n=43	B, n=41
Zakres	5,90 - 107,40	5,90 - 74,42	7,17 - 25,97	7,17 - 14,77	27,04 - 107,40	27,04 - 61,58
Średnia ± SD	36,94 ± 14,09	36,73 ± 13,57	11,99 ± 4,17	11,22 ± 2,51	42,75 ± 13,50	40,40 ± 7,49
Przedział ufności ±95%	35,43 - 38,46	35,27 - 38,20	9,98 - 14,01	9,97 - 12,47	38,60 - 46,91	38,04 - 42,77
Mediana	40,34	40,28	11,39	10,95	42,29	41,39
90 Percentyl	50,94	50,76	14,77	14,57	47,89	47,69
97,5 Percentyl	54,52	54,48	25,97	14,77	74,42	48,49
Skośność	-0,16	-0,56	2,11	-0,01	2,89	-0,02
Kurtoza	1,22	-0,34	6,58	-1,24	12,43	0,29

A – zbiór danych wyjściowych, B – zbiór danych po eliminacji wartości odstających, mg/kg

5. Podsumowanie

Przeprowadzone badania wykazały konieczność dokładnej analizy uzyskanych wyników przed przeprowadzeniem właściwej analizy chemometrycznej. Jest to istotne, szczególnie w przypadku występowania w zbiorze danych wartości znacznie różniących się od innych, tzw. wartości odbiegających. Wartości te, szczególnie w zbiorach o małej liczebności, mogą znacząco wpływać na charakter rozkładu uzyskanych wyników oraz powodować uzyskanie nieprawdziwego obrazu badanego obiektu. Wynik testu wskazujący, że wartość odbiegająca jest naprawdę większa niż oczekiwana względem posiadanego zbioru danych, powinien prowadzić do wnikliwej analizy postępowania przygotowawczego, tj. pobierania, przechowywania, preparatyki próbki oraz sposobu oznaczenia wartości badanego parametru w celu wykluczenia ewentualnych błędów popełnionych na którymś z tych etapów i w razie potrzeby powtórzenia oznaczenia.

LITERATURA

1. EPA 1998. Guidance for Data Quality Assessment, Practical Methods for Data Analysis, EPA QA/G-9, QA97 Version, EPA/600/R-96/084, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.
2. Mazerski J.: Podstawy chemometrii. Wydawnictwo Politechniki Gdańskiej, Gdańsk 2000.
3. EPA 1996. Guidance for Data Quality Assessment, Practical Methods for Data Analysis, EPA QA/G-9, QA96 Version, EPA/600/R-96/084, U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

Recenzent: Dr hab. inż. Barbara Białecka