

EFEKTYWNOŚĆ WYBRANYCH METOD SZACOWANIA UDZIAŁU ŹRÓDEŁ PUNKTOWYCH W ZANIECZYSZCZENIU WÓD POWIERZCHNIOWYCH

Marzena Marcinowska, Marek Nawalany
Instytut Systemów Inżynierii Środowiska, Politechnika Warszawska

1 WSTĘP

Przedmiotem pracy jest teoretyczna analiza przydatności oraz porównanie ze sobą wybranych metod statystycznych i heurystycznych, nazywanych w literaturze metodami chemometrycznymi, służących do szacowania wpływu punktowych źródeł zanieczyszczeń na jakość wód powierzchniowych. W pracy rozpatrywano następujące metody: analizę regresji wielu zmiennych, model CMB, analizę składowych głównych, model UNMIX, model PMF. Oceny poszczególnych metod dokonywano na podstawie danych pozyskanych z numerycznego modelu jakości wody stanowiącego tzw. rzeczywistość wirtualną. Wykonano eksperymenty służące określeniu udziałów trzech punktowych strumieni masy sześciu substancji w przyjętym punkcie kontrolnym.

1.1 Sformułowanie problemu

Występujące powszechnie zjawisko zanieczyszczenia środowiska wodnego przez użytkowników wód rodzi pytanie o to, kto i w jakim stopniu powinien ponosić koszty związane z przeciwdziałaniem degradacji ekosystemów, usunięciem zanieczyszczenia i likwidacją ewentualnych szkód. Powstaje problem rozkładu odpowiedzialności sprowadzający się do złożonego pytania: czy i w jakim stopniu koszty te powinno całkowicie lub częściowo ponosić społeczeństwo w postaci podatków, a w jakim stopniu za zanieczyszczenie wód powinien być odpowiedzialny podmiot zanieczyszczający [1].

Nałożenie na podmioty zanieczyszczające odpowiedzialności związanej z ich negatywnym

oddziaływaniem na środowisko i obciążenie ich kosztami związanymi z usuwaniem szkód jest jednym ze sposobów wymuszenia zachowań polegających na zapobieganiu zanieczyszczeniu środowiska. Tego typu podejście leży u podstaw *zasady: zanieczyszczający płaci*.

Obowiązujące w Polsce Prawo ochrony środowiska [2] stanowi, iż „kto powoduje zanieczyszczenie środowiska, ponosi koszty usunięcia skutków tego zanieczyszczenia”, natomiast „kto może spowodować zanieczyszczenie środowiska, ponosi koszty zapobiegania temu zanieczyszczeniu”. Prawo ochrony środowiska nie precyzuje jednak sposobu wskazania podmiotu zanieczyszczającego. Nie wypowiedza się, jak dzielona jest odpowiedzialność w przypadku zanieczyszczenia wód przez kilka podmiotów. Należy przypuszczać, iż uszczegółowienie tych dwóch zagadnień będzie przedmiotem odpowiednich Rozporządzeń Ministra Środowiska. Niniejsza praca antycypuje konieczność takiego uszczegółowienia. Nie stawia sobie jednak zadania ostatecznego rozstrzygnięcia tych zagadnień.

2 CEL I ZAKRES PRACY

Przedmiotem pracy jest teoretyczna analiza przydatności oraz porównanie ze sobą wybranych metod statystycznych i heurystycznych, nazywanych w literaturze *metodami chemometrycznymi*, służących do szacowania udziału źródeł punktowych w zanieczyszczeniu wody w okolicach ujęcia zlokalizowanego w zbiorniku wodnym.

Dane, na podstawie których określano efektywności poszczególnych metod chemometrycznych do szacowania udziału źródeł punktowych w ogólnym zanieczyszczeniu wody, pochodzą z wyników symulacji numerycznych modelu jakości wody Zbiornika Zegrzyńskiego traktowanego jako rzeczywistość wirtualna. Opis narzędzi matematycznych służących do zbudowania modelu hydrodynamiki oraz modelu jakości wody przedstawiono odpowiednio w rozdziale 4.1. Sposób pozyskiwania danych z modelu zawarto w rozdziale 4.2. Następnie, w rozdziale 5.1, opisano szereg wykonanych eksperymentów numerycznych. Na podstawie danych pochodzących z różnych eksperymentów numerycznych wyliczono, według koncepcji przedstawionej w rozdziale 5.2, rzeczywiste udziały źródeł punktowych. Do wyliczonych w ten sposób udziałów odnoszono następnie wyniki uzyskiwane przy wykorzystaniu wybranych metod chemometrycznych przedstawionych w rozdziale 5.3.

2.1 Oznaczenia i terminologia stosowana w pracy

Oznaczenia skrótów:

- CMB – *chemical mass balance*, model zbudowany w oparciu o chemiczny bilans masy,
- PMF – *positive matrix factorization*, dodatnia faktoryzacja macierzy,
- PCA – *principal component analysis*, analiza składowych głównych,
- PC – *principal component*, składowa główna, oś główna,
- FA – *factor analysis*, analiza czynnikowa
- SVD – *singular value decomposition*, dekompozycja do wektorów własnych,
- MLR(A) – *multivariate linear regression (analysis)* wielowymiarowa regresja liniowa,
- PCR – *principal component regression*, regresja oparta o składowe główne,
- PLS(M) – *partial least squares (modelling)*, model cząstkowych najmniejszych kwadratów,
- CA – *cluster analysis*, analiza skupień,
- LDA – *linear discriminant analysis*, analiza dyskryminacji liniowej,
- QDA – *quadratic discriminant analysis*, analiza dyskryminacji kwadratowej.

Oznaczenia symboli:

- N – liczba kroków pomiarowych,
- M – liczba substancji,
- K – liczba źródeł zanieczyszczeń, liczba istotnych czynników/składowych głównych,
- $\mathbf{P}_{N \times M}$ – macierz ładunków o wymiarach $N \times M$,
- \mathbf{Z} – skalowana macierz ładunków,
- \mathbf{M} – diagonalna macierz wykorzystywana do wyznaczenia normalizowanej macierzy ładunków,
- \mathbf{T} – macierz rotacji, macierz transformacji,
- \mathbf{K} – macierz kowariancji/korelacji,
- \mathbf{S} – macierz wartości czynnikowych,
- \mathbf{L} – macierz ładunków czynnikowych,
- \mathbf{E} – macierz błędów,
- $\mathbf{F}_{K \times M}$ – macierz profili źródeł,
- \mathbf{w} – wektor wag,
- \mathbf{S} – macierz udziałów źródeł,
- $\|\mathbf{A}\|_F$ – miara Frobeniusa dowolnej macierzy \mathbf{A} równa $\sqrt{\sum_n \sum_m A_{nm}^2}$.

W pracy w miarę możliwości stosowano ujednolicone oznaczenia, jakkolwiek nie uniknięto w kilku przypadkach stosowania tych samych symboli w innym znaczeniu niż przedstawiono powyżej.

Definicje wybranych pojęć:

- profil źródła* – względny bądź bezwzględny udział mas poszczególnych substancji w ogólnej masie zanieczyszczeń emitowanych przez źródło, często przedstawiany w postaci wykresu radarowego,
- frakcja* – procentowy udział masy danej substancji w ogólnej masie zanieczyszczeń,
- receptor* – miejsce, w którym badany jest udział/wpływ źródeł punktowych,
- udział źródła* – część masy wszystkich zanieczyszczeń emitowanych przez źródło docierająca do miejsca receptora, często przedstawiana w procentach,
- waga źródła* – udział procentowy masy wszystkich zanieczyszczeń pochodzących z danego źródła w masie zanieczyszczeń obserwowanych w miejscu receptora.

3 METODY WYKRYWANIA ŹRÓDEŁ ZANIECZYSZCZEŃ

Do szacowania wpływu różnych źródeł zanieczyszczeń wykorzystywano następujące modele:

- wielowymiarowa analiza regresji oraz analiza regresji oparta na składnikach głównych,
- model CMB (ang. *chemical mass balance*),
- analiza składowych głównych, analiza czynnikowa, dekompozycja macierzy do wektorów własnych,
- model UNMIX,
- dodatnia faktoryzacja macierzy PMF (ang. *positive matrix factorization*).

Punktem wyjścia do wszystkich analiz wykonywanych w pracy jest znajomość ładunków różnych substancji (\mathbf{P}) docierających do wyróżnionego miejsca (receptora), np. w rozważanym przypadku do ujęcia wody. W wyniku obliczeń uzyskuje się wagi (\mathbf{w}) nadane poszczególnym źródłom punktowym odpowiednio do ich udziału w ogólnym zanieczyszczeniu rejestrowanym w receptorze [3, 4, 5, 6, 7, 8, 9].

Źródła zanieczyszczeń mające wpływ na jakość wody w punkcie receptora pozostawiają charakterystyczny dla każdego z nich ślad (ang. *fingerprints*) w analizowanych próbkach wody. Udział poszczególnych źródeł w zanieczyszczeniu może być oszacowany, jeżeli znany jest chemiczny ślad każdego ze źródeł [10].

Dla ustalonej liczby źródeł K ładunek substancji m (P_m) w miejscu receptora jest sumą ładunków pochodzących z K źródeł. Jeżeli frakcja masy zanieczyszczenia substancją m w źródle k równa jest f_{mk} , natomiast S_k jest ogólną masą wszystkich zanieczyszczeń pochodzących ze źródła k , wówczas równanie bilansu masy dla substancji m może być zapisane w następujący sposób:

$$P_m = \sum_{k=1}^K f_{mk} S_k, \quad m = 1, \dots, M \quad (1)$$

Interpretacja fizyczna tego równania stanowi podstawę wszystkich modeli receptorowych. Wartości S_k są poszukiwanymi wielkościami udziałów źródeł w ogólnej masie zanieczyszczeń.

4 OBIEKT BADAŃ

4.1 Model obiektu

Obiektem rzeczywistym, na podstawie modelu którego analizowana była efektywność metod szacowania udziału źródeł punktowych w zanieczyszczeniu wody w okolicach ujęcia był Zbiornik Zegrzyński. Zbiornik ten powstał w latach 1958-1963 w wyniku spiętrzenia wód Narwi w miejscowości Dębe.

Do celów niniejszej pracy zbudowano model matematyczny Zbiornika Zegrzyńskiego, wykorzystując wyniki wcześniejszych prac badawczych prowadzonych w Instytucie Systemów Inżynierii Środowiska [11].

Model obiektu zbudowano w oparciu o francuski system TELEMAT [12]. W skład systemu wchodzi zestaw współpracujących ze sobą programów: SINUSX, MATISSE, STBTEL, TELEMAT2D, SUBIEF2D, RUBENS [13].

Szczegółowy opis metod numerycznych wykorzystywanych przez system TELEMAT można znaleźć w publikacji Hervouet'a i Van Haren'a [14, 15].

Wspomnieć należy, iż symulacje transportu zanieczyszczeń w modelu SUBIEF2D zostały przeprowadzone dla zmiennych, które nie ulegają zmianom podczas transportu w wodzie. Do takich substancji należeć mogą np. siarczan, chlorki, sól, potas oraz ewentualnie magnez i wapń. Tylko na podstawie takich parametrów, zachowujących bilans masy w zbiorniku, można szacować udziały punktowych źródeł zanieczyszczeń. W niniejszej pracy wszelkie modelowanie oraz obliczenia wykonywano dla substancji symbolicznych, bez przypisywania im konkretnych nazw i właściwości chemicznych. Dobór odpowiednich substancji do tego typu analiz dla rzeczywistego przypadku powinien być przedmiotem odrębnych badań nie wchodzących w zakres pracy. Nie podjęto także modelowania transportu substancji podlegających reakcjom chemicznym. Zagadnienie takie stanowi wciąż wyzwanie dla nauki, zwłaszcza w skali tutaj rozważanej.

4.2 Sposób pozyskiwania danych z modelu

Model obiektu posłużył do wykonania szeregu eksperymentów numerycznych i wygenerowania danych z miejsca receptora niezbędnych do dalszych analiz. Wykonano eksperymenty służące

określeniu udziałów trzech punktowych strumieni masy sześciu substancji w przyjętym miejscu.

Miejscem receptora w rozpatrywanym przypadku było ujęcie wody. Wyliczenia ładunku zanieczyszczeń dopływających do ujęcia symbolizowanego przez wybrany przekrój elementarny dokonano za pomocą utworzonego do tego celu makra w programie *STATISTICA*. Wszystkie wyniki modelu SUBIEF2D odczytywane przez postprocesor RUBENS [16] zapisywano w postaci plików ASCII możliwych do zaimportowania i dalszej obróbki w programie *STATISTICA* [17]. Pliki ASCII zawierające dla każdego z węzłów siatki następujące dane: współrzędną X, współrzędną Y, poziom dna zbiornika, poziom zwierciadła wody, składową poziomą i składową pionową wektora prędkości oraz wartości stężeń modelowanych substancji dla wszystkich chwil czasowych importowane były za pomocą makra utworzonego w języku *STATISTICA Visual Basic (SVB)* do jednego pliku zapisywanego w formacie programu *STATISTICA*. Następnie, przy wykorzystaniu innego makra *SVB*, z pliku tego wydobywano informacje dotyczące przebiegów czasowych wszystkich wymienionych zmiennych w węzłach wchodzących w skład interesujących przekrojów. Informacje te zapisywano w oddzielnych plikach wykorzystywanych następnie do wyliczeń ładunku.

Ładunek zanieczyszczeń P [g/s] przepływający przez pionową powierzchnię przekroju o szerokości $(b-a)$ oraz głębokości g przedstawia się równaniem:

$$\rho = \int_a^b V_n(x) \cdot C(x) \cdot g(x) dx \quad (2)$$

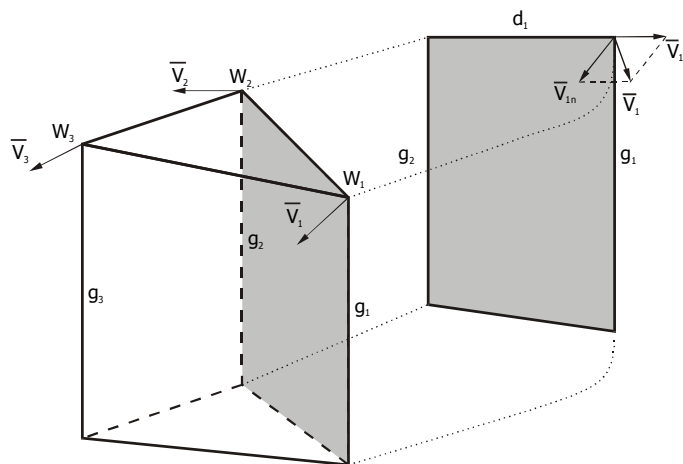
gdzie:

$V_n(x)$ – składowa prędkości normalna do powierzchni w punkcie x przekroju [m/s],

$C(x)$ – stężenie substancji w danym punkcie x przekroju poprzecznego [g/m³],

$g(x)$ – głębokość w punkcie x przekroju poprzecznego [m].

Do obliczeń ładunku P [g/s] zanieczyszczeń przepływających przez przekrój przedstawiony na rys. 1 wykorzystano metodę całkowania numerycznego Newtona-Cotes'a drugiego stopnia [18].



Rys. 1. Element siatki.

Ładunek P [g/s] przepływający przez przekrój oparty na krawędzi W_1W_2 obliczano według wzoru:

$$P = \frac{d_1}{6} \cdot \left(v_{1n} \cdot g_1 \cdot c_1 + 4 \cdot \frac{(v_{1n} + v_{2n})}{2} \cdot \frac{(g_1 + g_2)}{2} \cdot \frac{(c_1 + c_2)}{2} + v_{2n} \cdot g_2 \cdot c_2 \right)$$

gdzie:

d_1 – długość wektora $\overline{W_1W_2}$ [m],

v_{1n} – składowa normalna do krawędzi W_1W_2 z prędkości w węźle W_1 [m/s],

v_{2n} – składowa normalna do krawędzi W_1W_2 z prędkości w węźle W_2 [m/s],

c_1 – stężenie substancji w węźle W_1 [g/m³],

c_2 – stężenie substancji w węźle W_2 [g/m³],

g_1 – głębokość w węźle W_1 [m],

g_2 – głębokość w węźle W_2 [m].

Dane dotyczące współrzędnych kartezjańskich kolejnych węzłów W_1 , W_2 , W_3 elementu siatki, dane dotyczące głębokości g_1 , g_2 , g_3 zbiornika w poszczególnych węzłach, wektory prędkości wody przyłożone w poszczególnych węzłach $(\overline{v_1}, \overline{v_2}, \overline{v_3})$, dane dotyczące stężeń substancji dostępne były w postaci plików zapisanych w formacie programu *STATISTICA*.

Przebiegi czasowe ładunku zanieczyszczeń przepływających przez przekrój symbolizujący ujęcie wody liczone według powyższego schematu stanowiły rzeczywistość wirtualną.

5 WYBÓR METODY

5.1 Opis eksperymentów numerycznych

Symulacje numeryczne przygotowano dla różnych warunków (Warunki1, Warunki2,...,

Warunki*). Różne warunki oznaczały zróżnicowanie pod względem stężeń substancji emitowanych przez źródła punktowe, pod względem stężeń substancji w wodach dopływających oraz stężeń początkowych na obszarze całego zbiornika. Dla każdego z warunków przygotowywano szereg scenariuszy symulacji: case*3, case*11, case*12, case*13, case*0, gdzie * oznacza numer warunku. W scenariuszu pierwszym (case*3) modelowano sytuację, w której wszystkie trzy źródła emitowały zanieczyszczenia, w kolejnych trzech scenariuszach (case*11, case*12, case*13) uwzględniano emisję z tylko jednego głównego źródła, emisje z pozostałych dwóch źródeł były zaniedbywalne.

Scenariusz case*0 stanowiła symulacja przeprowadzona dla określenia rozchodzenia się

zanieczyszczeń pochodzących z dopływów do zbiornika, z zaniedbywalną emisją ze wszystkich trzech źródeł punktowych.

W kolejnych warunkach symulacji rozpatrywano te same profile źródeł, czyli udziały masy poszczególnych substancji emitowanych przez dane źródło w sumarycznej masie zanieczyszczeń pochodzących z tego źródła. Profile poszczególnych źródeł ustalono tak, aby wyraźnie się od siebie różniły. Dla każdego źródła charakterystyczne były dwie substancje i ich stężenie było dla danego źródła największe. W różnych warunkach różne były odchylenia standardowe profili. Schematyczny opis różnych warunków symulacji przedstawia tabela 1.

Tabela 1. Opis warunków symulacji.

Numer	Opis
Warunki1	Stałe stężenia na dopływach; zmienne stężenia w źródłach punktowych, zmienność stężeń wg rozkładu logarytmiczno-normalnego, profile źródeł punktowych według załączonych rysunków. <div style="display: flex; justify-content: space-around; margin-top: 10px;"> 1 źródło 2 źródło 3 źródło </div>
Warunki2	Zmienne stężenia na dopływach, średnie ładunki docierające z każdego z dopływów są sobie równe, zmienne stężenia w źródłach punktowych, średnie profile źródeł takie jak w przypadku Warunki1.
Warunki3	Założenia takie jak dla przypadku Warunki1, w emisjach źródeł występują przerwy.
Warunki4	Zmienne stężenia na dopływach, średnie ładunki docierające z każdego z dopływów są różne, jednakowe pozostają średnie wartości stężeń na dopływach; zmienne stężenia w źródłach punktowych, średnie profile jak wyżej, większe wartości odchylenia standardowego dla emisji substancji charakterystycznych. <div style="display: flex; justify-content: space-around; margin-top: 10px;"> </div>
Warunki5	Zmienne stężenia na dopływach, tak jak w przypadku Warunki4, średnie profile źródeł bez zmian, szeregi czasowe emisji ze źródeł o przebiegu sinusoidalnym.

Szeregi czasowe stężeń substancji emitowanych przez źródła generowane były za pomocą generatora liczb pseudolosowych (wyjątek stanowią symulacje według schematu Warunki5).

Stężenie substancji w konkretnej chwili czasowej reprezentowane było poprzez zmienną losową pochodzącą z rozkładu logarytmiczno-normalnego [19], zależną od przyjętych

parametrów rozkładu, tj. wartości średniej oraz współczynnika zmienności danych. W ten sposób były także generowane szeregi czasowe zanieczyszczeń dopływających do zbiornika. Zarówno w celu generowania liczb, jak i tworzenia plików wejściowych do programu napisano procedury w języku FORTRAN90.

5.2 Wyliczenia udziałów na podstawie danych modelowych

Symulacje przeprowadzane oddzielnie dla każdego punktowego źródła zanieczyszczeń (case*11, case*12, case*13) służyły wydobyciu informacji na temat wpływu zanieczyszczeń emitowanych przez poszczególne źródła na jakość wody w punkcie ujęcia. Symulacja dla wszystkich źródeł włączonych jednocześnie do modelu (case*33) służyła zbadaniu wypadkowego wpływu wszystkich źródeł. Symulacja z wyłączonymi źródłami punktowymi (case*0) miała na celu określenie tła zanieczyszczenia.

Szeregi czasowe ładunków zanieczyszczeń w miejscu ujęcia, pochodzące z kolejnych symulacji pozwoliły na oszacowanie udziałów zanieczyszczeń pochodzących z poszczególnych źródeł. Udział ten obliczono, wykorzystując uśrednione po czasie symulacji wartości ładunków wraz z wyliczonymi dla nich wartościami odchylenia standardowego.

Udział źródła k w całkowitej masie zanieczyszczeń obserwowanej w punkcie ujęcia, przy założeniu liniowości systemu oraz pasywności wskaźników, obliczono według wzoru:

$$s_k = \sum_{m=1}^M \bar{p}_m^k \quad (3)$$

natomiast wagę źródła według wzoru:

$$w_k = \frac{\sum_{m=1}^M \bar{p}_m^k}{\sum_{m=1}^M \bar{p}_m}, \text{ gdzie } \sum_{k=1}^K w_k = 1 \quad (4)$$

M – liczba substancji,

\bar{p}_m^k – ładunek m -tej substancji pochodzący z k -tego źródła [g/s],

\bar{p}_m – ładunek m -tej substancji pochodzący z K źródeł [g/s].

Zgodnie z teorią błędów [20] wartości błędów σ_f funkcji złożonych, takich jak np. \bar{w}_k obliczano według schematu:

$$\sigma_f = \sqrt{\sum_{i=1}^l \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_{x_i}^2} \quad (5)$$

gdzie:

l – liczba wielkości prostych tworzących funkcję f .

Zestawienie wyliczonych wag wraz z odpowiadającymi im wartościami błędów dla poszczególnych źródeł dla różnych warunków symulacji przedstawia tabela 2.

Tabela 2. Wagi poszczególnych źródeł w zanieczyszczeniu w miejscu ujęcia dla różnych warunków symulacji.

	Warunki 1	Warunki 2	Warunki 3	Warunki 4	Warunki 5
Źródło 1	0.54 ± 0.01	0.54 ± 0.01	0.55 ± 0.05	0.63 ± 0.06	0.66 ± 0.11
Źródło 2	0.23 ± 0.00	0.23 ± 0.01	0.22 ± 0.04	0.18 ± 0.03	0.21 ± 0.04
Źródło 3	0.23 ± 0.01	0.23 ± 0.01	0.23 ± 0.04	0.19 ± 0.03	0.13 ± 0.02

Wyliczone w przedstawiony sposób wagi traktowano jako wagi rzeczywiste, w odniesieniu do których określano następnie efektywność wybranych metod chemometrycznych.

5.3 Wyliczenia udziałów na podstawie wybranych metod chemometrycznych

Do wyliczeń udziałów źródeł punktowych w zanieczyszczeniu obserwowanym w miejscu receptora, czyli w miejscu ujęcia wody, wykorzystano następujące metody i modele:

- wielowymiarową regresję liniową,
- model CMB8,
- analizę składowych głównych,
- model UNMIX,
- model PMF2.

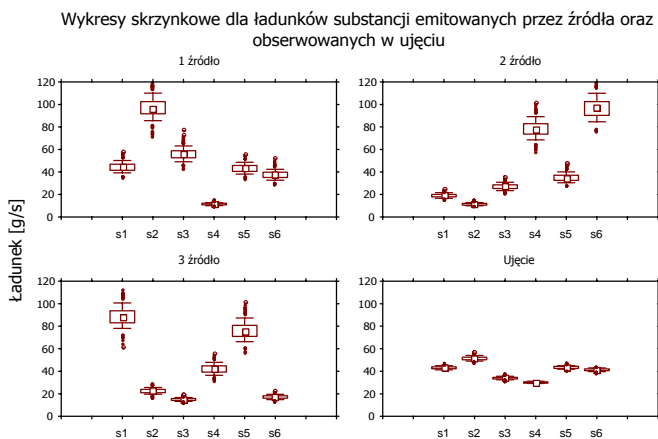
5.3.1 Przygotowanie danych

Na podstawie wyników pochodzących z symulacji case*3, pomniejszonych o odpowiednie dla każdego z rozpatrywanych warunków wartości zanieczyszczenia docierające wraz z wodami rzek dopływających, przygotowywano ciągi czasowe różniące się czasem próbkowania i wynoszącym w kolejnych szeregach odpowiednio 2,

4, 6 oraz 12 godzin. Uzyskiwano w ten sposób zbiory zawierające po 654, 327, 218 oraz 109 wyników pomiarów dla każdej z sześciu rozpatrywanych substancji, tworzące macierze danych wejściowych o wymiarach 654×6 , 327×6 , 218×6 oraz 109×6 .

5.3.2 Wstępne analizy statystyczne

Na rys. 2 przedstawiono wykresy skrzynkowe ładunków poszczególnych substancji emitowanych w źródłach punktowych oraz obserwowanych w miejscu ujęcia.



Rys. 2. Wykresy skrzynkowe ładunków substancji emitowanych przez źródła punktowe oraz ładunków substancji obserwowanych w ujęciu – Warunki 1.

W celu zweryfikowania, czy obserwacje pochodzące z trzech źródeł tworzą wyraźne różne skupienia w przestrzeni M -wymiarowej przeprowadzono analizę skupień. W przestrzeni tej badano położenie ładunków obserwowanych w miejscu ujęcia pochodzących z wyników symulacji case*3.

W tym celu tworzono jedną kolumnową macierz danych zawierającą w wierszach, dla każdego ze źródeł, obserwacje dotyczące emisji oraz obserwacje z miejsca ujęcia. Z tak zbudowanej macierzy danych wyodrębniano cztery skupienia przy wykorzystaniu grupowania metodą k -średnich.

5.3.3 Wielowymiarowa regresja liniowa

W celu zastosowania metody wielowymiarowej regresji liniowej, oprócz obserwacji w miejscu receptora (macierz $\mathbf{P}_{N \times M}$), wykorzystano wiedzę na temat emisji ze źródeł (macierz $\mathbf{F}_{M \times K}$). Elementy macierzy \mathbf{F} to uśrednione po czasie

wartości emisji ładunków sześciu substancji ($M=6$) pochodzących z trzech rozpatrywanych źródeł ($K=3$). Model wielowymiarowej regresji liniowej przedstawiony jest wówczas równaniem $\mathbf{P}_{N \times M} = \mathbf{S}_{N \times K} \mathbf{F}_{K \times M}^T$, gdzie każdy element macierzy \mathbf{P} można przedstawić równaniem $p_{nm} = s_{n1}f_{1m} + s_{n2}f_{2m} + s_{n3}f_{3m}$. Rozwiązanie dla macierzy \mathbf{S} uzyskiwane jest poprzez odpowiedni iloczyn macierzy:

$$\mathbf{S} = \mathbf{P} \mathbf{F} (\mathbf{F}^T \mathbf{F})^{-1} \quad (6)$$

Na podstawie kolumn macierzy \mathbf{S} uzyskanych w wyniku rozwiązania równania (6) wyliczono następnie wartości średnie

$$\bar{s}_k = \frac{1}{n} \sum_{i=1}^n s_{nk} \quad (7)$$

stanowiące, jaka część emisji z poszczególnych źródeł trafia do miejsca ujęcia.

Na podstawie wartości \bar{s}_k wyliczono wagi \bar{w}_k poszczególnych źródeł oraz odpowiadające im odchylenia standardowe σ_{wk} . Wagi wyliczono według wzoru $\bar{w}_k = \bar{s}_k \cdot \bar{P}_k / \bar{P}$, gdzie:

$$\bar{P} = \sum_{m=1}^M \frac{1}{N} \sum_{n=1}^N p_{nm}, \text{ natomiast } \bar{P}_k = \sum_{m=1}^M f_{km}.$$

Zgodnie z teorią błędów [20] wyliczono również wartości błędów σ_f funkcji złożonych, takich jak \bar{w}_k . Wartości wag wraz z wartościami błędów dla różnych warunków symulacji przedstawia tabela 3.

Tabela 3. Wagi źródeł dla różnych warunków symulacji wyliczone w oparciu o MLRA (dane co 2h).

	Warunki 1	Warunki 2	Warunki 3	Warunki 4	Warunki 5
Źródło 1	0.54 ± 0.04	0.53 ± 0.05	0.55 ± 0.09	0.61 ± 0.18	0.65 ± 0.13
Źródło 2	0.23 ± 0.02	0.24 ± 0.04	0.22 ± 0.05	0.13 ± 0.10	0.19 ± 0.06
Źródło 3	0.23 ± 0.02	0.23 ± 0.03	0.23 ± 0.05	0.26 ± 0.09	0.16 ± 0.02

5.3.4 Model CMB8

Do modelu CMB8 wprowadzono, dla każdego z warunków symulacji, obserwacje z miejsca ujęcia oraz względne profile źródeł wraz z estymatami błędów.

Estymaty udziałów źródeł oraz błędy estymacji uzyskiwano niezależnie dla każdego wiersza obserwacji, a następnie na podstawie wielokrotnych wyników wyliczono udziały

średnie. Dla każdego z warunków symulacji rozpatrywano ciągi o różnej długości, zawierające obserwacje prowadzone co 2, 4, 6 oraz 12h.

Tabela 4. Średnie wagi źródeł dla różnych warunków symulacji wyliczone przy wykorzystaniu modelu CMB8 (dne co 2h).

	Warunki 1	Warunki 2	Warunki 3	Warunki 4	Warunki 5
Źródło 1	0.54 ± 0.06	0.53 ± 0.06	0.55 ± 0.11	0.62 ± 0.16	0.64 ± 0.07
Źródło 2	0.23 ± 0.04	0.24 ± 0.04	0.22 ± 0.07	0.14 ± 0.09	0.20 ± 0.04
Źródło 3	0.23 ± 0.05	0.23 ± 0.05	0.23 ± 0.09	0.24 ± 0.11	0.16 ± 0.05

5.3.5 Analiza składowych głównych

Za pomocą analizy składowych głównych podjęto próbę wyodrębnienia liczby źródeł oraz ich profili wyłącznie na podstawie obserwacji w miejscu receptora.

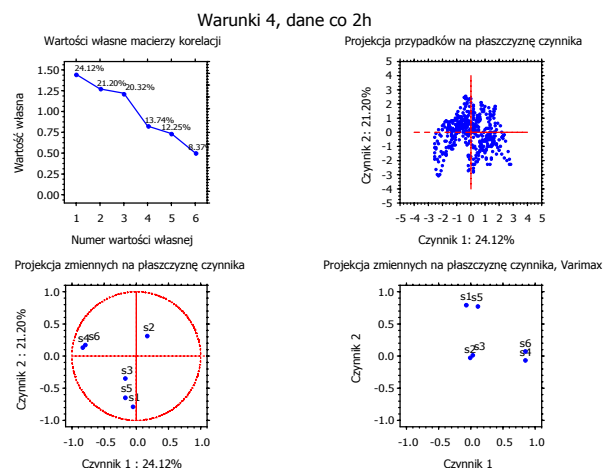
W tym celu, dla każdego z warunków, z obserwacji pochodzących z miejsca ujęcia tworzono kolumnową macierz danych $\mathbf{P}_{N \times M}$, którą następnie przy wykorzystaniu programu *STATISTICA* poddawano analizie składowych głównych. Zgodnie z algorytmem PCA dane wykorzystywane do analizy uprzednio autoskalowano. W wyniku PCA uzyskiwano macierz wartości czynnikowych oraz macierz ładunków czynnikowych. Macierze te następnie poddawano rotacji VARIMAX. Do wyodrębnienia istotnej liczby czynników stosowano kryterium jedności oraz wykres osuwiska.

Spośród wszystkich warunków symulacji wyodrębnienie składników (czynników) odpowiadających trzem źródłom punktowym obserwuje się najwyraźniej w przypadku Warunki4, po zastosowaniu rotacji VARIMAX. Wyniki te przedstawiono graficznie w postaci czterech wykresów: wykresu osuwiska, projekcji przypadków na płaszczyznę, projekcji zmiennych na płaszczyznę oraz projekcji zmiennych na płaszczyznę po zastosowaniu rotacji VARIMAX (rys. 3).

Wstępnie, na podstawie macierzy ładunków czynnikowych uzyskanych po rotacji VARIMAX, przeprowadzono przyporządkowanie czynników do poszczególnych źródeł.

Po zinterpretowaniu macierzy ładunków czynnikowych przystąpiono do analizy macierzy wartości czynnikowych. Dzięki przyporządkowaniu kolejnych czynników do poszczególnych źródeł zanieczyszczeń, kolumny macierzy

wartości czynnikowych traktowano jako odpowiadające szeregom czasowym sumarycznej masy zanieczyszczeń pochodzących od przyporządkowanych im źródeł. Następnie przystąpiono do analizy MLRA, traktując macierz obserwacji $\mathbf{P}_{N \times M}$ jako macierz zmiennych zależnych, natomiast macierz wartości czynnikowych $\mathbf{S}_{N \times K}$ jako macierz zmiennych niezależnych. Poszukiwano macierzy $\mathbf{L}_{K \times M}$. Na podstawie wartości elementów macierzy \mathbf{L} wyliczono względne (procentowe) profile źródeł.



Rys. 3. Wyniki analizy składowych głównych dla przypadku Warunki4 (dane co 2h).

Uzyskane macierze profili wykorzystywano jako macierze zmiennych niezależnych w kolejnej analizie MLRA, za pomocą której otrzymywano nowe macierze \mathbf{S}' zawierające w kolumnach szeregi czasowe sumarycznej masy zanieczyszczeń pochodzących z kolejnych źródeł. Na podstawie kolumn macierzy \mathbf{S}' wyliczono wartości średnie sumarycznej masy zanieczyszczeń. Uzyskane wyniki przedstawiano po przeliczeniu jako wagi poszczególnych źródeł w całkowitej masie zanieczyszczeń obserwowanych w miejscu ujęcia.

5.3.6 Model UNMIX

Za pomocą modelu UNMIX wykonano szereg analiz, opierając się na wynikach scenariuszy case*3 pochodzących z różnych warunków symulacji. Sprawdzano i porównywano rezultaty dla szeregów czasowych o różnym kroku próbkowania.

Tworzono odpowiednie macierze obserwacji. Ponadto dla schematu Warunki1 oraz Warunki5 rozpatrywano przypadek, w którym do macierzy

zawierających dane pochodzące z symulacji case*3 dołączano obserwacje z miejsca ujęcia pochodzące z symulacji case*221, case*223, case*231, czyli takich, na które wpływ jednego ze źródeł był znikomy.

Niestety, na podstawie większości z rozpatrywanych macierzy danych określenie udziałów poszczególnych źródeł za pomocą modelu UNMIX okazało się niewykonalne.

5.3.7 Model PMF2

Za pomocą modelu PMF2 wykonano szereg analiz w oparciu o wyniki scenariuszy case*3 pochodzące z różnych warunków symulacji.

Niejednoznaczność rozwiązań uzyskiwanych za pomocą modelu PMF2 próbowano zniwelować, wprowadzając wiedzę *a priori* o tym, iż niektóre elementy poszukiwanych macierzy reprezentujących profile źródeł są bliskie zeru. Było to możliwe poprzez zmianę parametru FPEAK. Rozpatrywano różne wartości tego parametru. Wagi źródeł dla różnych warunków symulacji i wartości parametru FPEAK= 0.05 przedstawia tabela 5.

Tabela 5. Średnie wagi źródeł dla różnych warunków symulacji wyliczone przy wykorzystaniu modelu PMF2 (FPEAK=0.05) (dane co 2h).

	Warunki 1	Warunki 2	Warunki 3	Warunki 4	Warunki 5
1czynnik	0.30 ± 0.02	0.34 ± 0.04	0.32 ± 0.05	0.30 ± 0.13	0.27 ± 0.09
2czynnik	0.36 ± 0.02	0.36 ± 0.04	0.33 ± 0.07	0.40 ± 0.13	0.39 ± 0.10
3czynnik	0.34 ± 0.02	0.30 ± 0.04	0.35 ± 0.06	0.31 ± 0.09	0.34 ± 0.06

5.4 Porównanie efektywności wybranych metod chemometrycznych

Porównanie efektywności wybranych metod chemometrycznych wykonano niezależnie dla dwóch grup metod. Do pierwszej grupy zaliczono wielowymiarową analizę regresji oraz model CMB8, czyli te metody, w algorytmie których niezbędna była wiedza *a priori* na temat profili źródeł mających wpływ na obserwacje w miejscu receptora. W grupie tej porównywano tylko efektywność w wyznaczaniu udziałów/wag źródeł w ładunku zanieczyszczeń w punkcie receptorowym. Do drugiej grupy zaliczono PCA, UNMIX, PMF, czyli te metody, w algorytmie których wykorzystywano tylko

wiedzę na temat obserwacji w receptorze. W grupie tych metod, oprócz efektywności w wyznaczaniu wag źródeł, porównywano również efektywność w wyznaczaniu profili źródeł.

W celu ilościowego porównania wybranych metod chemometrycznych dla każdej z nich wyznaczono współczynnik efektywności w wyznaczaniu wag oraz współczynnik efektywności w wyznaczaniu profili (dla metody PCA, UNMIX, PMF). Za współczynnik efektywności metody przyjęto pierwiastek z sumy kwadratów względnych różnic pomiędzy wagami (bądź profilami), wyliczonymi na podstawie tej metody, a wagami (bądź profilami) rzeczywistymi, wynikającymi z modelu jakości wody. Zestawienie współczynników efektywności wyliczonych dla wag przedstawia tabela 6. Przyjęto, iż graniczna wartość współczynnika efektywności dla wag wynosi 0.520, co oznacza, iż różnice względne pomiędzy wagami uzyskanymi za pomocą analizowanej metody a wagami rzeczywistymi wynoszą 30%. Metody, dla których współczynnik efektywności przyjmował wartości większe od wartości granicznej, uznano za nieefektywne dla rozpatrywanych warunków symulacji.

Tabela 6. Porównanie efektywności wybranych metod chemometrycznych w wyznaczaniu wag źródeł (dane co 2h).

Metoda	Warunki 1	Warunki 2	Warunki 3	Warunki 4	Warunki 5
MLRA	0.000	0.047	0.000	0.463	0.253
CMB8	0.000	0.047	0.000	0.345	0.238
PCA			0.337	1.162	
UNMIX	0.424		0.165		0.678
PMF2	1.010	0.451	0.815	0.727	1.024

Zestawienie współczynników efektywności dla profili źródeł wyznaczanych w oparciu o wybrane metody dla różnych warunków symulacji przedstawia tabela 7. Przyjęto, iż graniczna wartość współczynnika efektywności dla profili wynosi 1.273, co oznacza, iż różnice względne pomiędzy profilami wyznaczonymi za pomocą poszczególnych metod a profilem rzeczywistym wynoszą 30%. Metody, dla których współczynnik efektywności przyjmował wartości większe od wartości granicznej, uznano za nieefektywne dla rozpatrywanych warunków symulacji.

Tabela 7. Porównanie efektywności wybranych metod chemometrycznych w wyznaczaniu profili źródeł (dane co 2h).

Metoda	Warunki 1	Warunki 2	Warunki 3	Warunki 4	Warunki 5
PCA			1.39	4.25	
UNMIX	1.39		5.20		3.32
PMF2 (0.15)	7.92	8.42	1.64	7.66	7.61

Współczynniki efektywności dla wag wskazują na to, iż spośród pierwszej grupy metod lepsze rezultaty uzyskuje się w wyniku zastosowania modelu CMB8 niż w wyniku zastosowania analizy MLRA. Różnice pomiędzy tymi metodami są szczególnie wyraźne w przypadku danych wejściowych uzyskanych z bardziej skomplikowanego schematu symulacji (Warunki4).

Spośród metod zaliczonych do drugiej grupy pod względem efektywności w wyznaczaniu wag wyróżnia się model PCA oraz model UNMIX, a w przypadku Warunki2 także model PMF2. Po przeanalizowaniu współczynników efektywności dla profili (tabela 7) najbardziej efektywną metodą pozwalającą na wyznaczenie wag źródeł oraz na odwzorowanie profili tychże źródeł pozostaje model UNMIX, jednakże jedynie w przypadku Warunki1. W przypadku tym dodatkowo do modelu UNMIX wprowadzono dane będące punktami krawędziowymi. Poprzez wprowadzenie dodatkowych punktów krawędziowych w przypadku Warunki5, czyli bardziej skomplikowanego schematu symulacji, nie uzyskano już satysfakcjonujących współczynników efektywności. Współczynniki efektywności poniżej wartości granicznych zarówno dla wag, jak i dla profili uzyskano dla modelu PCA, ale jedynie w przypadku Warunki3.

6 WNIOSKI

◆ Wskazanie efektywnej metody służącej do szacowania wpływu źródeł punktowych na zanieczyszczenie wód powierzchniowych jest zadaniem niezwykle istotnym i, jak wykazano w pracy, jednocześnie bardzo skomplikowanym. Stopień komplikacji i niejednoznaczności rozwiązań występował mimo przyjęcia wielu uproszczeń, dalekich niejednokrotnie od warunków rzeczywistych. Niewątpliwie, w przypadku rzeczywistym, najważniejszym elementem umożliwiającym

w ogóle rozważanie zastosowania rozpatrywanych metod jest dobór odpowiednich substancji chemicznych jako substancji charakterystycznych emitowanych z punktowych źródeł zanieczyszczeń. Przede wszystkim substancje charakterystyczne powinny pojawiać się w punkcie receptorowym w postaci niezmienionej z pewnym opóźnieniem wynikającym z lokalizacji źródła zanieczyszczeń. Substancje te powinny wykazywać się znaczną trwałością w środowisku wód powierzchniowych, czyli nie powinny podlegać jednostkowym procesom fizycznym, chemicznym lub biochemicznym, które mogą wpływać na zmiany stężeń. Ponadto wybrane związki powinny być trudno przyswajalne przez organizmy zasiedlające rozpatrywane środowisko, co uniemożliwiłoby ich kumulację w poszczególnych ogniwach łańcucha zależności troficznych. W pracy przyjęto założenie, że rozpatrywane substancje spełniają wymienione wcześniej założenia, dzięki czemu zagwarantowano zachowanie bilansu ich masy w zbiorniku.

- ◆ Gdy dostępna jest pełna wiedza na temat profili emisji ze wszystkich źródeł zanieczyszczeń mających wpływ na jakość ujmowanej wody, wówczas satysfakcjonujące rezultaty na temat udziałów źródeł uzyskać można poprzez zastosowanie modelu CMB8, lub też o wiele prostszego modelu wielowymiarowej regresji liniowej. W sytuacji rzeczywistej można podjąć próby zaadaptowania wymienionych modeli dla konkretnych szczegółowo zinwentaryzowanych obiektów.
- ◆ W niektórych sytuacjach bardzo przydatną i dającą wysoce zadowalające wyniki jest analiza składowych głównych. Mimo, iż jest to metoda czysto matematyczna, możliwe dzięki niej staje się wyłonienie czynników odpowiadających źródłom mającym istotny wpływ na obserwacje pochodzące z miejsca receptora. Interpretacja fizyczna czynników staje się zdecydowanie łatwiejsza po zastosowaniu rotacji VARIMAX.

Wyniki PCA są przejrzyste i możliwe do zinterpretowania w zagadnieniach związanych z detekcją źródeł punktowych tylko i wyłącznie w pewnych szczególnych przypadkach, w których to także ogromne znaczenie ma wiedza *a priori* o źródłach zanieczyszczeń mogących mieć potencjalny

wpływ na obserwacje w miejscu ujęcia. Sposób wyodrębniania składowych głównych za pomocą metody PCA opiera się na założeniu ortogonalności kolumn macierzy ładunków czynnikowych, a w rzeczywistości rzadko zdarza się, iż profile źródeł są ortogonalne.

- ◆ Obiecujące wydawało się być zastosowanie modelu UNMIX do zagadnień szacowania udziałów źródeł w zanieczyszczeniu wód powierzchniowych. Jednakże warunkiem koniecznym do znalezienia prawidłowego rozwiązania za pomocą tego modelu jest występowanie w analizowanym zbiorze danych pewnych obserwacji szczególnych. W sytuacji braku takich obserwacji naruszone zostają założenia modelu UNMIX i rozwiązania, choć możliwe do uzyskania, są dalekie od rzeczywistości.
- ◆ Ze znaczną rozważką należy również traktować wyniki uzyskiwane z modelu PMF2. Ograniczenia co do nieujemności elementów macierzy wynikowych modelu PMF2 nie są w stanie zredukować problemu niejednoznaczności rozwiązań. Stosowanie różnych sposobów rotacji czasami przybliża do rozwiązania rzeczywistego, jednakże nie ma metody określającej słuszność zastosowanej rotacji. Model PMF jest ciągle doskonalony i być może część niejednoznaczności jego rozwiązań zostanie wyeliminowana.
- ◆ Spośród przeanalizowanych metod nie można wskazać żadnej, za pomocą której możliwe byłoby zidentyfikowanie źródeł zanieczyszczeń poprzez trafne wyznaczenie ich profili tylko i wyłącznie na podstawie obserwacji w miejscu ujęcia. Jak wynika z przedstawionych analiz, do efektywnego wyznaczania wag/udziałów źródeł w zanieczyszczeniu w miejscu receptora konieczna jest zazwyczaj pewna wiedza *a priori* o profilach chemicznych źródeł zanieczyszczeń.

- 4) Henry R.C. (1997) History and fundamentals of multivariate air quality receptor models, Chemometrics and Intelligent Laboratory Systems, Elsevier.
- 5) Park E.S., Henry R.C., Spiegelman C.H. (1999) Determining the number of major pollution sources in Multivariate Air Quality Models, NRCSE Technical Report Series.
- 6) Alloway B.J., Ayres D.C. (1999) Chemiczne podstawy zanieczyszczenia środowiska, Wydawnictwo Naukowe PWN.
- 7) Hemond H. F., Fecher E.J. (1994) Chemical Fate and Transport in the Environment, Academic Press, San Diego.
- 8) Lu R.S. et al. (1999) Analysis of reservoir water quality using fuzzy synthetic evaluation, Stochastic Environmental Research and Risk Assessment, Springer-Verlag.
- 9) Pesce S.F. et al. (2000) Use of water quality indices to verify the impact of Cordoba city (Argentina) on Suquia river, Water Research, Elsevier.
- 10) Henry R.C. (2002) Receptor modelling, Encyclopedia of Environmentrics, John Wiley&Sons.
- 11) Filocha M. (1997) Model hydrodynamiki Zbiornika Zegrzyńskiego, Praca Dyplomowa, Politechnika Warszawska.
- 12) Hervouet J. M., Lang P. (1994) TELEMAT-2D Version 3.0 – User Manual, Electricité de France
- 13) Quiquempoix O., Boulet T., Hervouet J.M. (1998) Matisse Software, User Manual, Electricité de France.
- 14) Hervouet J.M., Haren L. (1994) TELEMAT-2D Version 3.0 – Principle note, Electricité de France.
- 15) Filocha M., Nawalany M. (1998) Model hydrodynamiki Zbiornika Zegrzyńskiego, Prace Naukowe Politechniki Warszawskiej.
- 16) Graphical post-processor RUBENS – User Manual Version 4.1, Electricité de France.
- 17) STATISTICA 6.0 (2001) StatSoft Polska.
- 18) Gerald C.F., Wheatley P.O. (1999) Applied Numerical Analysis, Addison-Wesley.
- 19) McBean E.A., Rovers F.A. (1998) Statistical procedures for analysis of environmental monitoring data&risk assessment, Prentice Hall PTR.
- 20) Taylor J.R. (1995) Wstęp do analizy błędu pomiarowego, Wydawnictwo Naukowe PWN.

BIBLIOGRAFIA

- 1) Loga-Karpińska M., Nawalany M. (1997) Controlling polluters of water reservoirs using integrated monitoring and modelling systems, Operational Water Management, Refsgaard&Karalis.
- 2) Prawo Ochrony Środowiska Dz.U.2001.62. 627.
- 3) Hopke P.K. (1991) Receptor Modelling for Air Quality Management, Elsevier.