Akademia Górniczo-Hutnicza im. St. Staszica
Wydział Geologii, Geofizyki i Ochrony Środowiska

Ewa Kmiecik

ROZPRAWA DOKTORSKA

Optymalizacja gęstości opróbowania sieci monitoringowych jakości wód podziemnych

Promotor:
prof. dr hab. inż. Jadwiga Szczepańska

Kraków, czerwiec 2001

Składam serdeczne podziękowania
Pani Profesor ***Jadwidze Szczepańskiej***
za pomoc okazaną mi w trakcie pisania
niniejszej pracy.

Dziękuję również Prezesowi SPSS Polska sp. z o.o.
Piotrowi Komornickiemu
za udostępnienie programu Neural Connection.

Spis treści

Cel i zakres pracy

Kilka lat temu w sprawozdaniu dotyczącym przeglądu systemów informacji o stanie środowiska w krajach Europy Środkowej (m.in. w Polsce, Węgrzech i ówczesnej Czechosłowacji) podkreślano fakt, że na wszystkich poziomach systemów największą słabością jest brak wszechstronnego i kompleksowego systemu informacji, powiązanego przestrzennie i czasowo (Wiatr, 1998). Z przedstawionych danych wynikało, że w naszym kraju 90% danych o środowisku jest zbierane poprzez ankiety, a tylko 10% przez zastosowanie monitoringu — główną przyczyną takiego stanu rzeczy są, niestety, względy finansowe. One również decydują o tym, że systemy kontroli jakości/zapewnienia jakości QA/QC, które obligatoryjnie winny być stosowane w dużych sieciach monitoringowych (Nielsen, 1991; Witczak, Adamczyk, 1994; Szczepańska, Kmiecik, 1998) nie zawsze są stosowane w praktyce.

„Monitoring wód podziemnych jest kontrolno-decyzyjnym systemem oceny dynamiki antropogenicznych przemian tych wód. Polega on na prowadzeniu w wybranych, charakterystycznych punktach (punktach obserwacyjnych, posterunkach, stacjach) powtarzalnych pomiarów i badań stanu zwierciadła wód podziemnych i ich jakości a także interpretacji ich wyników w aspekcie ochrony środowiska wodnego" (Kleczkowski [Ed.], 1997).

Coraz większe znaczenie wód podziemnych, bardzo często jedynego źródła wód pitnych o dobrej jakości, stanowiło przyczynę podjęcia w Polsce badań jakości tych wód w trzech rodzajach sieci monitoringowych: krajowej, regionalnych i lokalnych (Hordejuk, 1993; Hordejuk, Gawin, 1994; Kropka, Różkowski, 1994; Witczak et al., 1994a,b; Prażak et al., 1996; Witkowski, 1997; Kazimierski, Sadurski, 1999).

Monitoring jakości wód podziemnych w naszym kraju prowadzony jest od ponad 10 lat. W bazach danych zgromadzono ogromne ilości obserwacji i pomiarów prowadzonych w sieciach lokalnych, regionalnych i w sieci krajowej.

Wyniki zgromadzone w bazach danych służą do oceny stanu jakości wód podziemnych oraz stopnia ich degradacji w układzie przestrzenno-czasowym. Na podstawie tych danych podejmowane są także decyzje o charakterze remediacyjnym lub ekonomicznym (finansowym — nakładanie kar) oraz odbywa się kompleksowe zarządzanie gospodarką wodną (Dyrektywa Unii Europejskiej 2000/60/EC). Dane te muszą więc cechować się wysokim stopniem pewności — stąd wynika konieczność ciągłej kontroli ich jakości.

Celem niniejszej rozprawy jest udowodnienie iż:

- poprzez analizę rozkładu badanych wskaźników fizyko-chemicznych wód za pomocą programów do statystycznej analizy danych: SPSS PL v. 10.0, QI Analyst 3.5 DB, ROB 2 i wyznaczenie odpowiednich parametrów kontroli jakości danych (DL, PDL, σ^2_{tech}) można dokonać weryfikacji danych pomiarowych z sieci monitoringowych jakości wód podziemnych;
- za pomocą sieci neuronowych (program Neural Connection) można prognozować jakość wód w nieopróbowanym punkcie monitoringowym, w oparciu o przeprowadzone badania wód podziemnych w punktach sąsiednich (zagadnienia **predykcji**);

- na podstawie tych prognoz można optymalizować gęstość opróbowania sieci monitoringowej jakości wód podziemnych, co pozwoli na obniżenie kosztów badań (opróbowanie i analiza), bez wpływu na poziom wiarygodności obserwowanych trendów zmian jakości w układzie przestrzennym;
- za pomocą sieci neuronowych, w oparciu o wartości wskaźników fizyko-chemicznych wód w danym punkcie monitoringowym można dokonać **klasyfikacji** tego punktu do obszaru o określonym zagospodarowaniu terenu (zagadnienia klasyfikacji).

Eksperymenty komputerowe w zakresie weryfikacji danych oraz optymalizacji gęstości opróbowania zostały wykonane przez autorkę na badaniach przeprowadzonych w latach 1993–1994, w sieci regionalnego monitoringu jakości wód podziemnych dorzecza górnej Wisły (Witczak et al., 1993a,b,c,d; Witczak et al., 1994; Witkowski, 1997),

Dane (oznaczenia 55 wskaźników fizyko-chemicznych wód) poddano weryfikacji na trzy sposoby: wyznaczono granice oznaczalności badanych wskaźników (laboratoryjną DL i praktyczną PDL), oszacowano udział wariancji technicznej σ^2_{tech} w wariancji całkowitej σ^2_{tot} na podstawie wyników badań próbek dublowanych, z wykorzystaniem klasycznej analizy wariancji ANOVA oraz elastycznego postępowania statystycznego (*robust statistics*) oraz dokonano analizy rozkładu tych wskaźników (rozdz. 1).

Wyłączono z dalszej analizy obserwacje anomalne, obarczone błędami grubymi, oznaczenia wskaźników fizyko-chemicznych cechujące się niską precyzją, i oznaczenia tych wskaźników, w których ponad 20% stanowiły wyniki poniżej granicy oznaczalności DL.

Dane literaturowe wskazują, że o jakości uzyskanych wyników pomiarów badanych wskaźników fizyko-chemicznych wód decydują głównie: proces opróbowania (Nielsen, 1991), precyzja zastosowanej metody analitycznej i warunki, w jakich wykonywane są oznaczenia — powtarzalność i odtwarzalność pomiarów (Międzynar. Słownik, 1996; Huber, 1997). To skłoniło autorkę do podjęcia dodatkowych badań, których obiektem były wody podziemne występujące w wapieniach górnej jury, w obszarze miasta Krakowa (artezyjski Zdrój Królewski). Do szczegółowych rozważań wybrano wyniki oznaczeń cynku (przedstawiciela metali ciężkich), który z uwagi na łatwość migracji powszechnie występuje w wodach podziemnych (Macioszczyk, 1987).

W ramach doskonalenia procesu opróbowania próbki ze Zdroju Królewskiego pobierane były w latach 1998–2000 sprzętem jednorazowego użytku firmy Millipore (37 próbek), zamiast sprzętu do opróbowania wielokrotnego użytku (Eijkelkamp) stosowanego w monitoringu jakości wód podziemnych dorzecza górnej Wisły w latach 1993–1994 (20 próbek). Do oznaczania cynku w wodach podziemnych wykorzystano, w miejsce stosowanej w RMWP dorzecza górnej Wisły metody AAS, metodę analityczną o niższej granicy oznaczalności — ICP-AES. Te dwa czynniki, tj. metodyka opróbowania i metodyka oznaczeń analitycznych danego wskaźnika oraz wykonywanie badań w warunkach powtarzalności i odtwarzalności wywierają decydujący wpływ na precyzję oznaczeń, a zatem na wyniki, na podstawie których ocenia się stan jakości wód podziemnych na danym obszarze (rozdz. 1).

Zweryfikowaną bazę danych dla sieci RMWP dorzecza górnej Wisły (16 wskaźników fizyko-chemicznych o dostatecznej precyzji spośród 55 analizowanych) można było dopiero wykorzystać do prognozowania zmian jakości wód w układzie przestrzennym i optymalizacji gęstości sieci opróbowania za pomocą sieci neuronowych (rozdz. 3).

Przygotowano trzy warianty danych zweryfikowanych w oparciu o wymienione wyżej parametry, różniące się liczbą zmiennych (wskaźników fizyko-chemicznych) i obserwacji (punktów RMWP):

- zbiór zawierający wszystkie zweryfikowane wskaźniki fizyko-chemiczne (16) i punkty monitoringowe o klasach zagrożenia wód AB, C, D (167 punktów RMWP);
- zbiór zawierający wszystkie zweryfikowane wskaźniki fizyko-chemiczne (16), ale punkty monitoringowe ograniczone do klasy zagrożenia AB (151 punktów RMWP);

- zbiór zawierający punkty monitoringowe o klasie zagrożenia AB i 6 wskaźników zweryfikowanych (są to wskaźniki, w których wystąpiła najmniejsza liczba braków danych, $n \leqslant 5$).

Na podstawie tych danych przeprowadzono eksperymenty prognozowania jakości wód w punkcie monitoringowym o określonych współrzędnych w oparciu o dane dla punktów sąsiednich oraz klasyfikacji punktu monitoringowego (na podstawie wyników oznaczeń wskaźników fizyko-chemicznych w tym punkcie) do obszaru o określonym użytkowaniu terenu.

Różne warianty danych wejściowych umożliwiły ocenę wpływu na jakość uzyskiwanych prognoz liczby zweryfikowanych wskaźników fizyko-chemicznych oraz liczby punktów monitoringowych w bazie danych wejściowych.

Do rozwiązania zagadnień predykcji i klasyfikacji jakości wód podziemnych w układzie przestrzennym wykorzystano modele sieci neuronowych z grupy sieci nadzorowanych (MLP, RBF, Bayesa). Eksperymenty polegały na budowaniu różnych modeli sieci, zmianie ich parametrów i analizie uzyskiwanych wyników prognoz (rozdz. 3).

Modele sieci neuronowych budowano i testowano w programie Neural Connection (SPSS, 1997, 1999), udostępnionym dla celów niniejszej rozprawy doktorskiej, dzięki uprzejmości prezesa firmy **SPSS Polska sp. z o.o.**, **Piotra Komornickiego**.

Najlepsze wyniki prognoz wskaźników fizyko-chemicznych wód na podstawie współrzędnych punktu monitoringowego — najmniejsze błędy względne prognoz — uzyskano dla sieci RBF. Dla takiego modelu przeprowadzono następnie próbę optymalizacji gęstości opróbowania sieci monitoringowej. Przeprowadzone symulacje wykazały, że poprzez ograniczenie liczby opróbowanych punktów, np. o ok. 15%, można uzyskać wiarygodne informacje o stanie jakości wód na danym obszarze, co w praktyce pozwoli znacznie zmniejszyć nakłady finansowe i ograniczyć czas trwania badań monitoringowych.

Pliki z analizowanymi danymi, pliki z modelami budowanych sieci neuronowych oraz pliki wynikowe i raporty z przeprowadzonych analiz znajdują się na płycie CD-ROM dołączonej do niniejszej pracy. Zamieszczono tam również pełny tekst pracy w formacie *.pdf.

rozdział 1

Regionalny monitoring jakości wód podziemnych RMWP dorzecza górnej Wisły

Do prognozowania zmian jakości wód podziemnych w układzie przestrzennym wykorzystano wyniki badań prowadzonych w roku 1993, w sieci Regionalnego Monitoringu Jakości Wód Podziemnych RMWP dorzecza górnej Wisły. Sieć ta składa się ze 172 punktów RMWP, z czego w obszarze RZGW Kraków znajduje się 117, zaś w obszarze RZGW Katowice — 55 punktów.

Analizie poddano wyniki badań jakości wód podziemnych pobranych w I serii opróbowania (okres mokry, V–IX 1993). W serii tej opróbowaniem i analizą objęto 167 punktów RMWP, gdyż punkty 11012, 21024, 21047, 21052 i 21060 (wg numeracji punktów w bazie MONBADA), ze względu na niezakończony proces ich adaptacji nie zostały opróbowane (Witczak et al., 1994).

1.1. Charakterystyka sieci

Regionalny monitoring jakości wód podziemnych RMWP obejmuje obszar dorzecza górnej Wisły od źródeł do ujścia rzeki Sanny (48 270 km^2). Na początku lat dziewięćdziesiątych obszar ten uznano za pilotowy do wprowadzenia w Polsce zintegrowanego zlewniowego zarządzania gospodarką wodną (Monitor Polski Nr 6 z 1991 r.). Program pilotowy objął obszar dwóch Regionalnych Zarządów Gospodarki Wodnej — RZGW Katowice i RZGW Kraków (rys. 1.1).

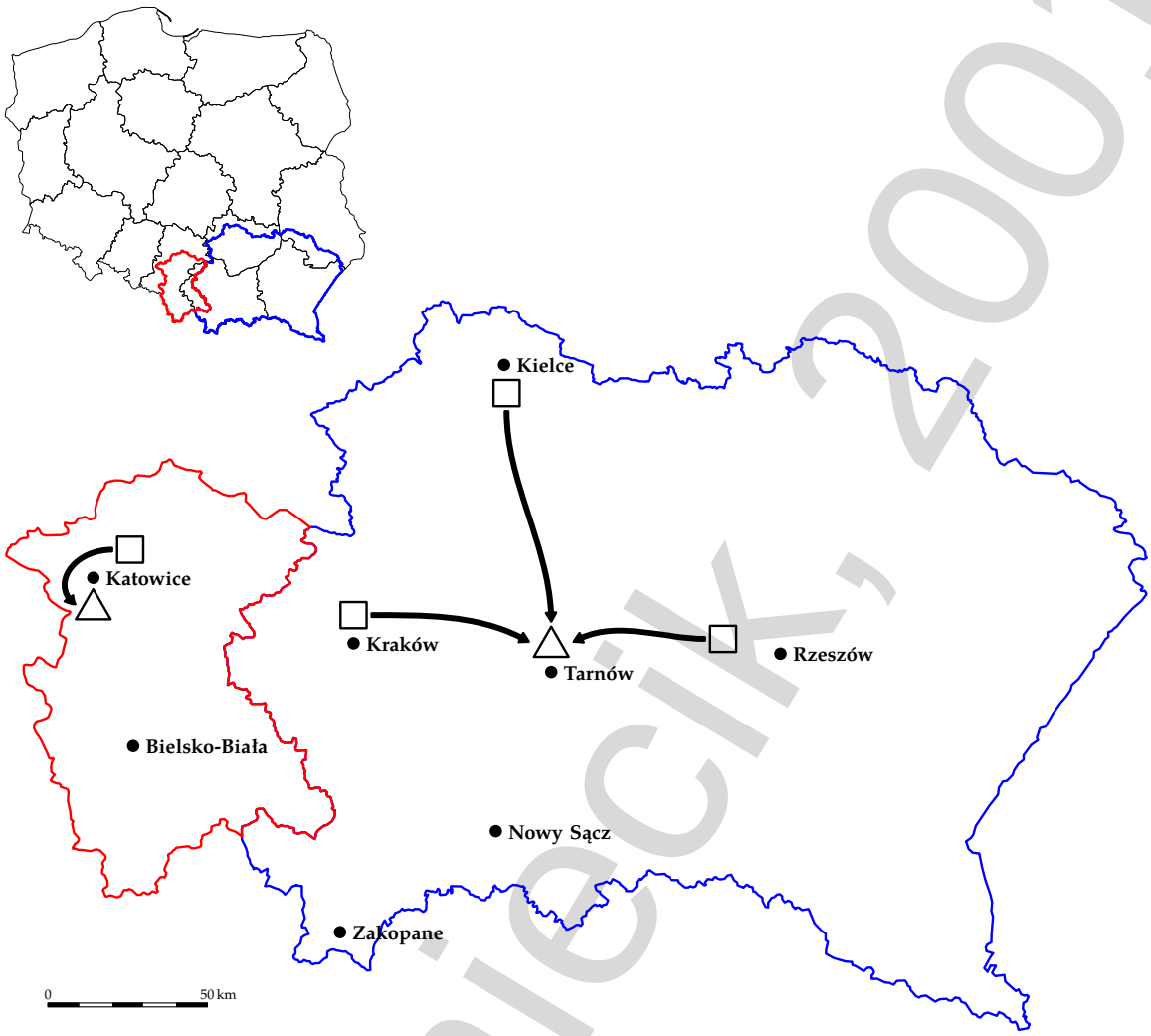
Program monitoringu regionalnego realizowany był przez zespoły: Katedry Hydrogeologii i Geologii Inżynierskiej Uniwersytetu Śląskiego (RZGW Katowice) oraz Zakładu Hydrogeologii i Ochrony Wód AGH w Krakowie we współpracy z Przedsiębiorstwem Geologicznym S.A. w Krakowie i Oddziałem Świętokrzyskim Państwowego Instytutu Geologicznego w Kielcach (RZGW Kraków). Badania składu chemicznego wód prowadzone były w dwóch laboratoriach środowiskowych: WIOŚ w Tarnowie (dla obszaru RZGW Kraków) i OBiKŚ w Katowicach (dla obszaru RZGW Katowice).

RMWP obejmuje obszar o bardzo zróżnicowanej morfologii, od gór typu alpejskiego (Tatry) po nizinną część w widłach Wisły i Sanu. Szczegółową charakterystykę obszaru badań można znaleźć w monografii (Dynowska, Maciejewski, 1991).

W obszarze zlewni górnej Wisły wydzielono ogółem 53 GZWP, z czego 14 występuje na obszarze RZGW Katowice, 32 GZWP na obszarze RZGW Kraków, zaś 7 GZWP przynależy do obu RZGW.

Wody podziemne występują w utworach czwartorzędu, trzeciorzędu, kredy, jury, triasu, karbonu i dewonu.

Wśród głównych zbiorników wód podziemnych (GZWP) dominują zbiorniki szczelinowo-porowe oraz szczelinowo-krasowe, mało odporne na zanieczyszczenia (Witczak et al., 1994; Prażak et al., 1996; Witkowski, 1997; Bednarczyk, 1998; Siwek, 1999). GZWP tworzą podstawową bazę eksploatacji wód podziemnych, stąd monitorowanie ich jakości stanowi niezwykle ważny problem.



Rysunek 1.1. Lokalizacja regionalnego monitoringu jakości wód podziemnych (RMWP) w zlewni górnej Wisły. Objaśnienia: granice RZGW Katowice —, granice RZGW Kraków —; □ — lokalizacja laboratoriów polowych opróbowania wód podziemnych; △ — laboratoria środowiskowe badania jakości wód podziemnych w Katowicach i Tarnowie; → — kierunki transportu próbek RMWP

Tabela 1.1. Charakterystyka ilościowa punktów sieci regionalnego monitoringu jakości wód podziemnych (RMWP) dorzecza górnej Wisły

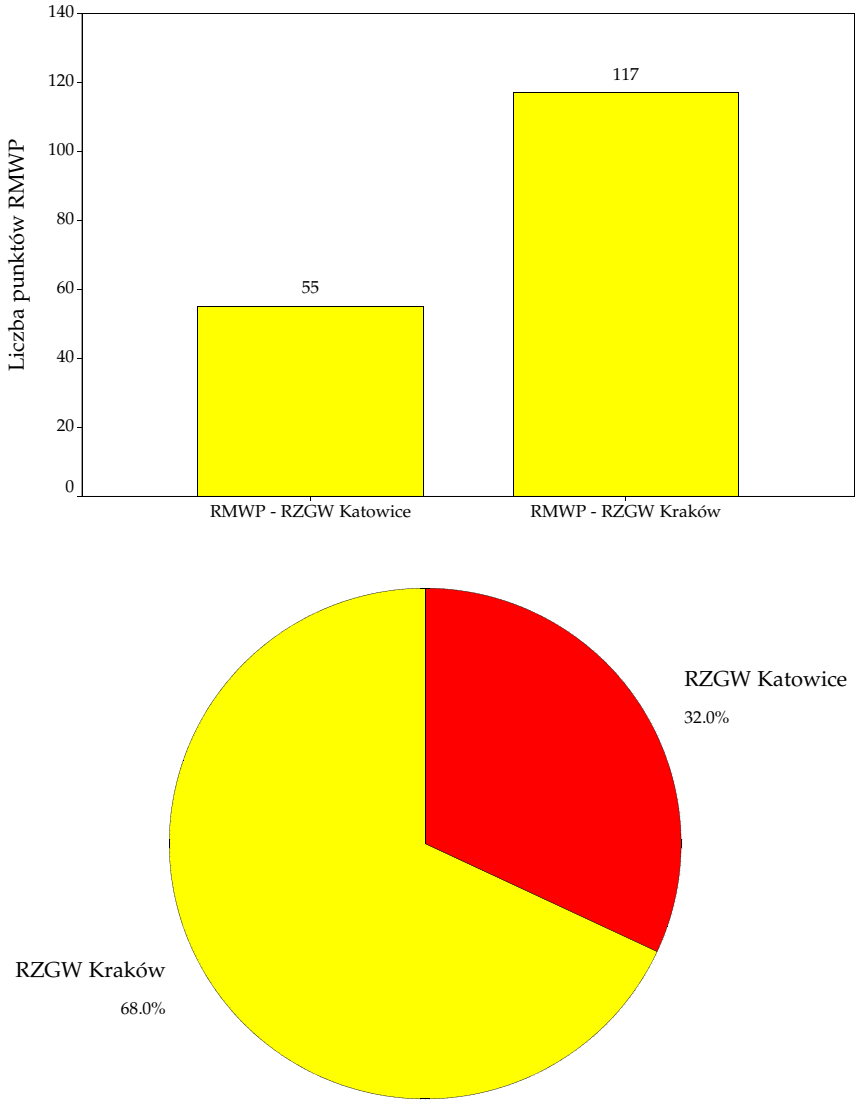
Obszar	Powierzchnia	Liczba punktów RMWP	Gęstość sieci RMWP
RZGW Katowice	7 380.2 km^2 (15.3%)	55 (32%)	1 pkt RMWP/134.2 km^2
RZGW Kraków	40 890.0 km^2 (84.7%)	117 (68%)	1 pkt RMWP/349.5 km^2
Razem:	48 270.0 km^2 (100.0%)	172 (100%)	1 pkt RMWP/280.6 km^2

Sieć regionalnego monitoringu jakości wód podziemnych w zlewni górnej Wisły obejmuje aktualnie 172 punkty, w tym 55 punktów znajduje się w obszarze RZGW Katowice a 117 RMWP w obszarze RZGW Kraków (rys. 1.2, tab. 1.1).

Struktura punktów sieci regionalnej jest następująca:

- 121 punktów stanowią studnie eksploatacyjne;
- 5 punktów, to otwory obserwacyjne;
- 3 punkty, to studnie kopane;
- 43 punkty — źródła.

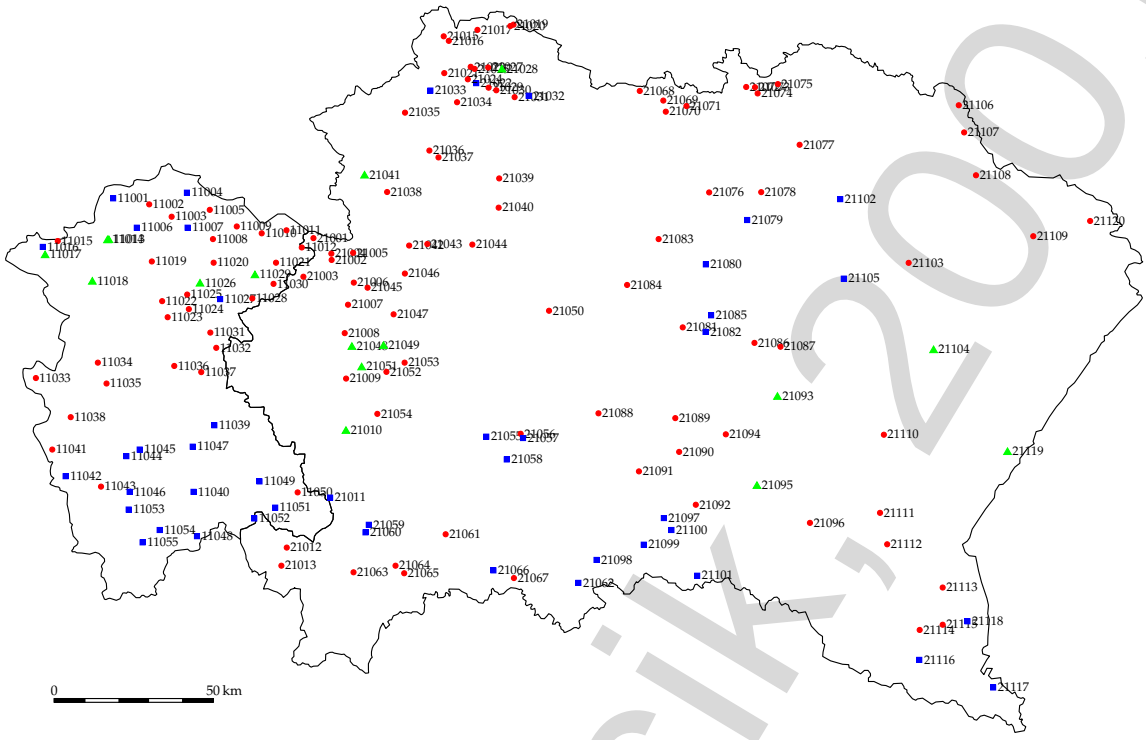
Szczegółowa charakterystyka punktów tworzących sieć RMWP została przedstawiona w wielu pracach (Witczak et al., 1994; Witkowski, 1997; Bednarczyk, 1998; Szczepańska, Kmiecik, 1998; Siwek 1999). Opis punktów wchodzących w skład sieci RMWP znajduje się w dodatku A, zaś ich lokalizację przedstawiono na rysunku A.1 (str. 149). Lokalizacja każdego punktu RMWP została tak dobrana, aby mógł on być reprezentatywny dla możliwie dużych obszarów, na które będzie można przenosić uzyskane wyniki badań monitoringowych. Z tego też względu na punkty RMWP wybierano przede wszystkim studnie eksploatacyjne i źródła zbierające wody z całego obszaru spływu a nie piezometry dające punktowe informacje.



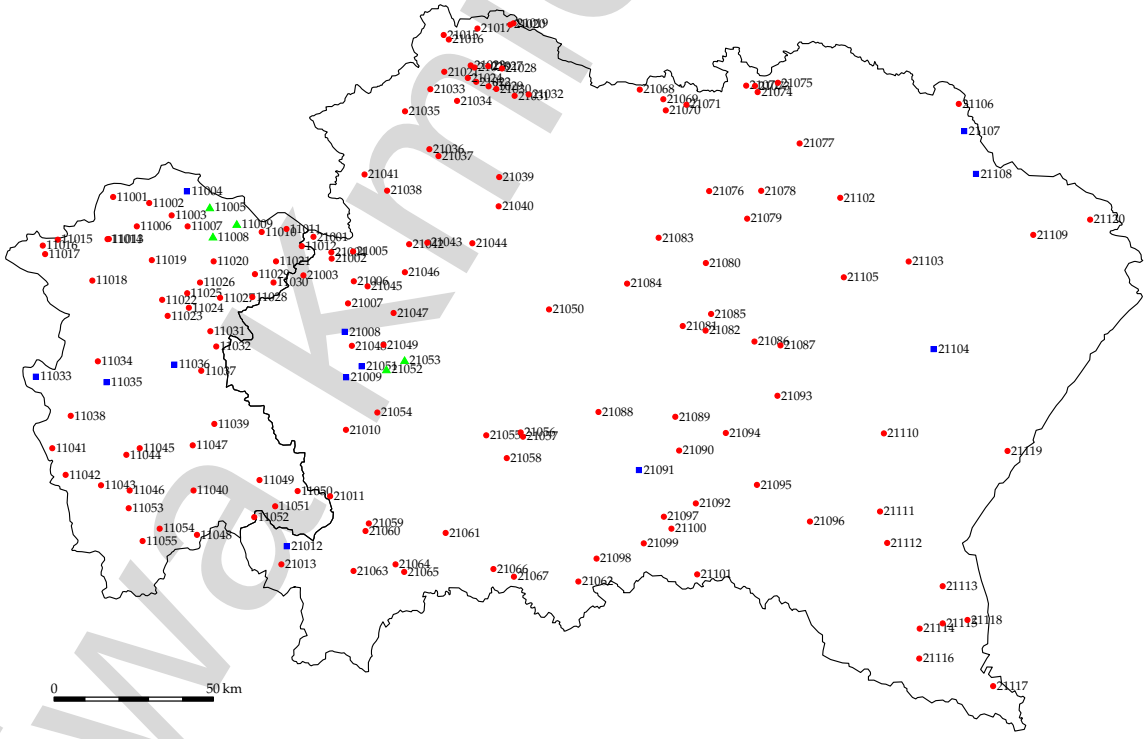
Rysunek 1.2. Charakterystyka ilościowa punktów tworzących sieć RMWP dorzecza górnej Wisły (przynależność do RZGW Katowice i RZGW Kraków)

Wszystkie punkty tworzące sieć RMWP spełniały następujące kryteria (Witczak et al., 1994):

- otwór badawczy (studnia) ujmuje tylko jedną warstwę wodonośną;
- w promieniu 2 km nie ma istotnego wpływu lokalnych ognisk zanieczyszczeń (obecność lokalnych ognisk zanieczyszczeń i rodzaj zagospodarowania terenu oceniono w promieniu 100, 500 i 2000 m od punktu RMWP);
- materiały użyte w konstrukcji studni są niereaktywne, nie powinny kontaminować wody;
- właściciel terenu, na którym znajduje się punkt RMWP wyraża zgodę na czynności związane z pompowaniem i opróbowaniem punktu RMWP.

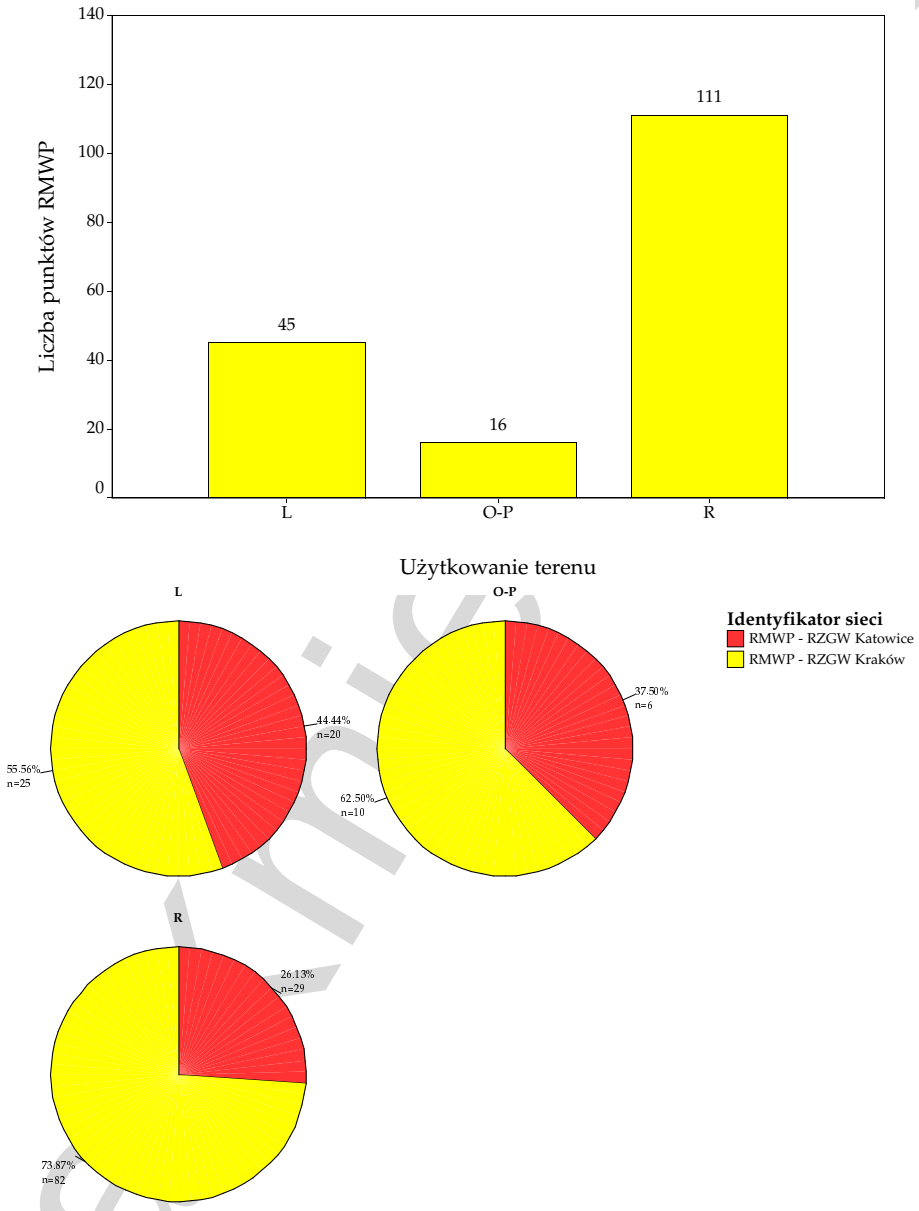


Rysunek 1.3. Regionalny monitoring jakości wód podziemnych (RMWP) w zlewni górnej Wisły, przynależność punktów monitoringowych do różnych obszarów zagospodarowania terenu (● — R (rolnicze); ■ — L (leśne); ▲ — O-P (osiedlowo-przemysłowe)). Numeracja punktów zgodna z ich identyfikacją w bazie MONBADA



Rysunek 1.4. Regionalny monitoring jakości wód podziemnych (RMWP) w zlewni górnej Wisły, przynależność punktów monitoringowych do różnych klas zagrożenia wód (● — klasa AB; ■ — klasa C; ▲ — klasa D). Numeracja punktów zgodna z ich identyfikacją w bazie MONBADA

Przy lokalizacji punktów RMWP uwzględniono ich reprezentatywność dla określonego typu antropopresji związanej z użytkowaniem terenu (rys. 1.3). Wydzielono trzy główne formy użytkowania terenu: rolnicze (R), leśne (L) i osiedlowo-przemysłowe (O-P), jako podstawę do oceny zagospodarowania terenu przyjmowano obszar o promieniu 500 m wokół punktu RMWP. Klasyfikacja punktów RMWP z obszarów RZGW Katowice i RZGW Kraków (172 RMWP) wykazała, iż są one rozmieszczone na obszarach o użytkowaniu: rolniczym — 65.0% punktów, leśnym — 25.5% punktów i osiedlowo-przemysłowym — 9.5% punktów (rys. 1.5).

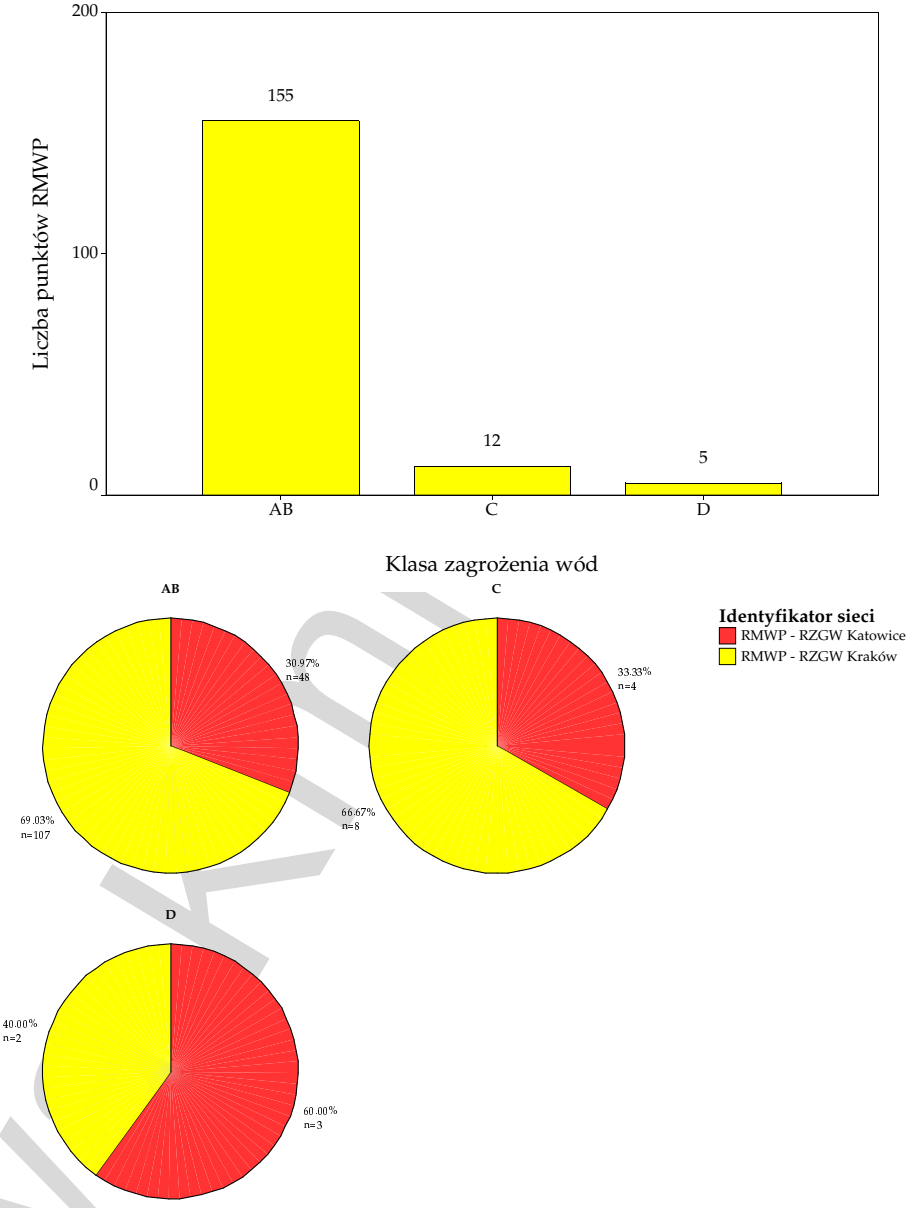


Rysunek 1.5. Przynależność punktów RMWP (RZGW Katowice i RZGW Kraków) do obszarów o różnym użytkowaniu terenu, według dokumentacji z lat 1993–94 (Witczak et al., 1994): L — leśne; O-P — osiedlowo-przemysłowe; R — rolnicze

Wartości te są zbliżone do form użytkowania ziemi w dorzeczu górnej Wisły, gdzie użytki rolne stanowią 58.4%, użytki leśne — 32.7% oraz inne — 8.9% (GRID, 1993). Taki dobór punktów umożliwia ocenę oddziaływania na jakość wód podziemnych dorzecza górnej Wisły obszarowych ognisk zanieczyszczeń związanych z typem zagospodarowania terenu oraz nakładających się na to — także obszarowych — zanieczyszczeń atmosfery.

Jako podstawowe kryterium dla oceny opóźnienia reakcji monitoringu na antropopresję przyjęty został czas migracji wody z powierzchni terenu do monitorowanej warstwy wodonośnej. Zastosowano klasyfikację zbliżoną do stosowanej dla oceny potencjalnego zagrożenia głównych zbiorników wód podziemnych GZWP (Kleczkowski [Ed.], 1990). W celu zapewnienia odpowiedniej liczebności obserwacji w poszczególnych klasach, ograniczono się do trzech klas, poprzez połączenie klas A i B w jedną AB (rys. 1.4, 1.6):

- **klasa AB** (czas migracji do 25 lat) — wody zagrożone;
- **klasa C** (czas migracji od 25 do 100 lat) — wody słabo zagrożone;
- **klasa D** (czas migracji ponad 100 lat) — wody praktycznie niezagrożone.



Rysunek 1.6. Przynależność punktów RMWP (RZGW Katowice i RZGW Kraków) do obszarów o różnej klasie zagrożenia wód podziemnych, według dokumentacji z lat 1993–94 (Witczak et al., 1994): AB — wody zagrożone; C — wody słabo zagrożone; D — wody praktycznie niezagrożone

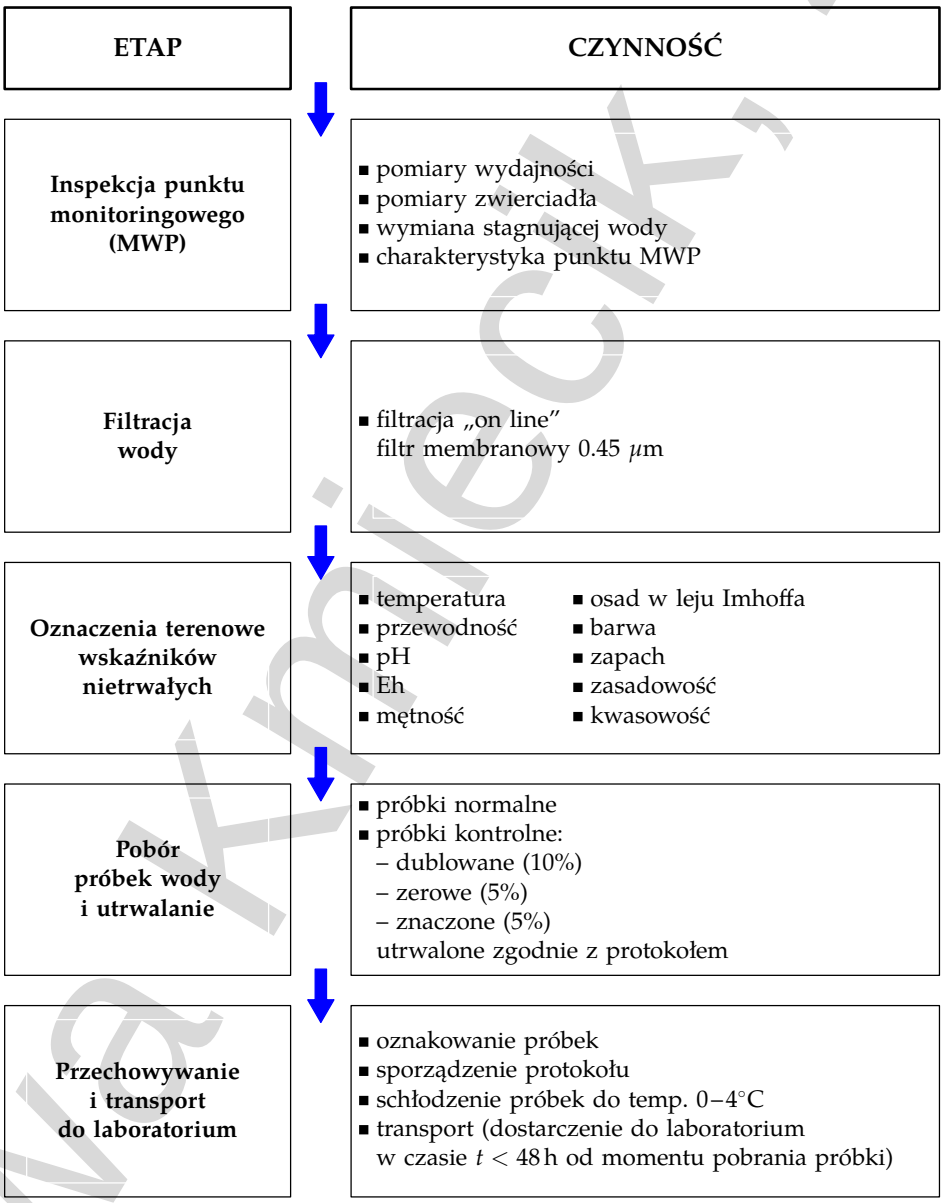
Wśród punktów tworzących sieć RMWP, zgodnie z przyjętymi założeniami, dominuje klasa AB — 90.1% (rys. 1.6). Punkty należące do tej klasy mają dostarczać informacji o stanie

zanieczyszczenia wód pod wpływem antropopresji. Punkty klas C (7.0%) i D (2.9%) mają informować o długoletnich zmianach jakości wód nie objętych jeszcze intensywną antropopresją.

1.1.1. Sprzęt, metodyka oraz zakres oznaczeń analitycznych

Opróbowanie sieci RMWP w roku 1993 zostało wykonane za pomocą sprzętu terenowego wielokrotnego użytku firmy Eijkelkamp (Witczak, Adamczyk, 1994), zainstalowanego w czterech samochodach-laboratoriach. Jedno ruchome laboratorium użytkowane było na obszarze RZGW Katowice, zaś trzy na obszarze RZGW Kraków (rys. 1.1).

Zakres czynności związanych z poborem próbek wód podziemnych, metodyką ich utrwalania oraz terenowym oznaczaniem niestabilnych cech i składników wody (10 wskaźników fizyko-chemicznych), przedstawiono schematycznie na rysunku 1.7.



Rysunek 1.7. Schemat opróbowania i obróbki próbek stosowany w monitoringu jakości wód podziemnych (wg Witczaka i Adamczyka, 1994)

Próbki filtrowano w terenie, bezpośrednio przy ich poborze z punktów RMWP przez filtr membranowy 0.45 μm. Filtrację prowadzono w systemie *on line*, bez styku wody z powietrzem.

Zakres analiz realizowano zgodnie z ZTE (Kleczkowski et al., 1991; Różkowski et al., 1991) oraz wytycznymi PIOŚ (Staniewicz-Dubois, 1991, 1995; Błaszyk, Macioszczyk, 1993). Analiza obejmowała składniki nietrwałe, oznaczane zgodnie z wytycznymi PIOŚ (Staniewicz-Dubois, 1991, 1995; Witczak, Adamczyk, 1994) obligatoryjnie w terenie przy poborze próbki wody oraz składniki nieorganiczne i organiczne, oznaczane w laboratorium.

Analiza terenowa obejmowała 10 wskaźników fizyko-chemicznych, oznaczanych przy poborze próbek wody przez grupy terenowe: temperatura, przewodność elektrolityczna właściwa, pH, potencjał utleniająco-redukcyjny Eh, mętność, osad w leju Imhoffa, barwa, zapach, zasadowość ogólna i mineralna, kwasowość ogólna.

Metodyka oznaczeń tych wskaźników została szczegółowo omówiona w „Katalogu wybranych wskaźników...", t. I (Witczak, Adamczyk, 1994) oraz w „Prawie ochrony środowiska Wspólnoty Europejskiej" (1996) a także zestawiona w tabeli 1.2.

Próbki wody pobrane z sieci monitoringowej były transportowane w ciągu 24–48 godzin do laboratorium: WIOŚ w Tarnowie dla obszaru RZGW Kraków i OBiKŚ Katowice dla obszaru RZGW Katowice. Oba laboratoria stosowały ten sam system zbierania i obróbki danych LIMS (*Laboratory Information Management System*).

Zakres pomiarów laboratoryjnych obejmował wskaźniki wchodzące, zgodnie z zaleceniami PIOŚ (Błaszyk, Macioszczyk, 1993), w zakres tzw. analizy szczegółowej. Obejmuje ona oprócz wskaźników podstawowych także wskaźniki dodatkowe.

Wskaźniki podstawowe (20 wskaźników) — substancje rozpuszczone, twardość ogólna, twardość węglanowa, agresywny CO_2, utlenialność (ChZT $KMnO_4$), krzemionka, azot amonowy, azot azotynowy i azot azotanowy, chlorki, siarczany, wodorowęglany (zasadowość), fosforany, sód, potas, wapń, magnez, żelazo, mangan, glin.

Wskaźniki dodatkowe (25 wskaźników) — fluorki, bor, metale ciężkie: cynk, miedź, ołów, kadm, nikiel, chrom, rtęć, współczynnik absorpcji UV, rozpuszczony węgiel organiczny (DOC), azot organiczny (Kjeldahla), substancje ropopochodne, fenole lotne, substancje powierzchniowo czynne anionowe, WWA (benzo-a-piren), chloroform, trójchloroetylen (trichloroeten), czterochloroetylen (tetrachloroeten) oraz pestycydy: DDT, DDE, DDD, gamma-HCH (lindan), metoksychlor (DMDT).

Razem analiza obejmowała 55 cech i składników wody, z czego 10 oznaczano w terenie a pozostałe w laboratorium. Badania składu chemicznego wód prowadzone były przez dwa laboratoria: WIOŚ Tarnów i OBiKŚ Katowice, przy zastosowaniu metod zalecanych dla potrzeb monitoringu jakości wód podziemnych (Witczak, Adamczyk, 1994, 1995; Prawo ochrony środowiska, 1996).

W tabeli 1.2 zestawiono wykorzystywane przez laboratoria metody analityczne wraz z granicami oznaczalności analizowanych wskaźników fizyko-chemicznych.

1.1.2. Program kontroli jakości

Sieć regionalnego monitoringu jakości wód podziemnych dorzecza górnej Wisły, ze względu na liczbę punktów monitoringowych (łącznie 172 punkty RMWP) objęta była specjalną procedurą kontroli jakości QA/QC badań laboratoryjnych i terenowych (Nielsen, 1991; Witczak et al., 1993; Witczak, Adamczyk 1994; Szczepańska, Kmiecik, 1998).

Program terenowy QA/QC w dorzeczu górnej Wisły (obejmującym obszary dwóch Regionalnych Zarządów Gospodarki Wodnej: w Katowicach i w Krakowie) polegał na pobraniu (tym samym sprzętem co próbki normalne) i analizie (w tym samym zakresie co normalne próbki wody) dodatkowych próbek kontrolnych. W pierwszej serii opróbowania sieci regionalnej (1993 rok) próbki te stanowiły ok. 22.6% próbek normalnych, z czego:

- 4.7% próbki zerowe — pobrane tym samym sprzętem co próbki normalne, ale z użyciem jako medium wody dejonizowanej. Odbywały one tę samą obróbkę, transport i przechowywanie jak próbki normalne, i były wykorzystane do wyznaczenia praktycznej granicy oznaczalności PDL (pobrano 8 próbek).

- 13.7% to próbki dublowane — pobierane z losowo wybranych punktów RMWP jako duplikaty próbek normalnych, służące do oceny precyzji oznaczeń (pobrano 23 próbki).
- 4.2% próbki znaczone o znanym składzie lub dodatku, umożliwiające ocenę dokładności oznaczeń — ze względu na brak reprezentatywnego materiału odniesienia (RM, CRM) badaniu poddano wodę o stabilnym składzie chemicznym i „wieku" wykluczającym oddziaływania antropogeniczne (samowypływ ze Zdroju Królewskiego w Krakowie; pobrano 7 próbek).

Wyniki programu kontroli jakości, przeprowadzonego w sieci RMWP dorzecza górnej Wisły oraz ocenę jakości badań analitycznych wykonywanych w trakcie opróbowania sieci RMWP w latach 1993–1994 zamieszczono m.in. w raporcie końcowym z badań (Witczak et al., 1994) oraz pracach: Witkowski, 1997; Bednarczyk 1998.

1.1.2.1. Granica oznaczalności DL i praktyczna granica oznaczalności PDL

Granica wykrywalności L_D (USP XXIII; Fleming et al., 1997) jest zdefiniowana jako najmniejsze stężenie analitu w próbce, które może być wykryte, niekoniecznie oznaczone w danych warunkach eksperymentalnych.

Granica oznaczalności L_Q (USP XXIII; Fleming et al., 1997) to najmniejsze stężenie analitu w próbce, które może być oznaczone z wymaganą dokładnością i precyzją, w danych warunkach eksperymentalnych.

W literaturze podawane są różne formuły na wyznaczanie granic wykrywalności L_D i oznaczalności L_Q (Szczepańska, Kmiecik, 1998), stąd wyniki obliczeń tych granic będą się różnić, w zależności od zastosowanego wzoru.

Zalecane do stosowania są wzory przedstawione w słowniku pojęć analitycznych („Terminology and Definitions"), publikowanym w czasopiśmie „Accreditation and Quality Assurance — Journal for Quality, Comparability and Reliability in Chemical Measurement".

Fleming et al. (1997) definiują tam granicę wykrywalności L_D jako najmniejsze stężenie analitu w próbce, które może być wykryte, niekoniecznie oznaczone (w danych warunkach eksperymentalnych).

Granicę tę (wyrażoną jako stężenie lub ilość), najmniejszy sygnał danej metody analitycznej, jaki może być wykryty z zadowalającą pewnością, uzyskuje się ze wzoru:

$$L_D = \bar{x}_l + k\sigma_l \tag{1.1}$$

gdzie: L_D — granica wykrywalności; $\bar{x}_l$ — wartość średnia z wyników oznaczeń próbek ślepych; σ_l — oszacowane odchylenie standardowe wyników oznaczeń próbek ślepych; k — numeryczny współczynnik odpowiadający żądanemu poziomowi ufności. Zaleca się prowadzić obliczenia dla $k = 3$, zatem:

$$L_D = \bar{x}_l + 3\sigma_l \tag{1.2}$$

Granica oznaczalności L_Q określona jest jako najmniejsze stężenie analitu w próbce, które może być dokładnie oznaczone (w danych warunkach eksperymentalnych).

Granicę tę (wyrażoną jako stężenie lub ilość) uzyskuje się ze wzoru:

$$L_Q = \mathrm{DL} = \bar{x}_l + k\sigma_l \tag{1.3}$$

gdzie: $L_Q = \mathrm{DL}$ — granica oznaczalności; $\bar{x}_l$ — wartość średnia z wyników oznaczeń próbek ślepych; σ_l — oszacowane odchylenie standardowe wyników oznaczeń próbek ślepych; k — numeryczny współczynnik odpowiadający żądanemu poziomowi ufności. Zaleca się prowadzenie obliczeń dla $k = 6$, czyli:

$$L_Q = \mathrm{DL} = \bar{x}_l + 6\sigma_l \tag{1.4}$$

Na podstawie wyników oznaczeń próbek zerowych, pobieranych w ramach terenowego programu kontroli jakości QA/QC, wyznacza się praktyczną granicę oznaczalności PDL. Stężenia analizowanych składników w tych próbkach nie powinny odbiegać od stężeń notowanych dla próbek ślepych przygotowywanych i analizowanych przez laboratorium w ramach programu laboratoryjnego QA/QC, przy zastosowaniu tej samej metodyki badań analitycznych. Stężenia w próbkach zerowych o wartościach wyższych od laboratoryjnej granicy oznaczalności DL są na ogół wynikiem kontaminacji tych próbek w trakcie poboru, utrwalania, transportu, itp.

Wśród najczęstszych przyczyn kontaminacji należy wymienić np. kurz, spaliny, przenoszenie zanieczyszczeń na używanym sprzęcie, możliwość ługowania określonych składników ze sprzętu i pojemników, niewłaściwie prowadzony proces filtracji próbek, a także odczynniki zastosowane do konserwacji próbek. Część z tych czynników może być wyeliminowana przez stosowanie odpowiednich pojemników i odczynników, pozostaje wówczas problem czystości opróbowania.

Wartość PDL można uzyskać na dwa sposoby:

- obliczeniowo — analogicznie jak laboratoryjną granicę oznaczalności, z tym, że zamiast próbek ślepych bada się próbki zerowe:

  $$\mathrm{PDL} = \bar{x}_{zer} + k\sigma_{zer} \tag{1.5}$$

  gdzie: PDL — to praktyczna granica oznaczalności; $\bar{x}_{zer}$ — wartość średnia z pomiarów próbek zerowych; σ_{zer} — oszacowane odchylenie standardowe; k — współczynnik wynikający z liczebności zbioru dla 95% przedziału ufności rozkładu normalnego „jednostronnego" (Szczepańska, Kmiecik, 1998). Fleming et al. (1997) zalecają wykonywanie obliczeń dla $k = 6$, stąd:

  $$\mathrm{PDL} = \bar{x}_{zer} + 6\sigma_{zer} \tag{1.6}$$

- graficznie — za pomocą siatki probabilistycznej; szczegóły dotyczące tej metody wyznaczania PDL można znaleźć w literaturze — Fresenius et al., 1988; Szczepańska, Kmiecik, 1998.

W ramach programu QA/QC prowadzonego w sieci RMWP dorzecza górnej Wisły wyznaczono praktyczne granice oznaczalności metodą obliczeniową bądź graficzną (Witczak et al., 1994; Bednarczyk, 1998).

Wartości PDL zestawiono w tabeli 1.3 wraz z granicami oznaczalności DL deklarowanymi przez laboratoria i maksymalnymi dopuszczalnymi stężeniami MDP analizowanych wskaźników w wodach pitnych (zgodnie z rozporządzeniem ministra zdrowia Dz.U. Nr 82, poz. 937 z 4 września 2000 roku oraz Dyrektywą Unii Europejskiej 98/83/EC z 3 listopada 1998 roku).(1) Praktyczne granice oznaczalności w zestawieniu z maksymalnymi dopuszczalnymi stężeniami analizowanych wskaźników w wodach pitnych wskazują na ile analizy laboratoryjne są w stanie określić stan jakości wód w sieci.

Pojęcie praktycznej granicy oznaczalności PDL wiąże się ściśle z precyzją badań hydrogeochemicznych. Wartość PDL ma znaczenie szacunkowe, bowiem informuje od jakiego stężenia można oczekiwać, że w warunkach rutynowych, w prawidłowo wyposażonym laboratorium uzyska się zadowalającą precyzję wyników.

Praktyczna granica oznaczalności PDL powinna mieć wartość jak najbliższą laboratoryjnej granicy oznaczalności DL; w idealnym przypadku PDL = DL (PDL/DL = 1).

Z tabeli 1.3 wynika, że nie będą na pewno wiarygodne wyniki oznaczeń rtęci, gdyż PDL/DL $\approx$ 18.5. Podobna sytuacja ma miejsce w przypadku oznaczeń chloroformu (PDL/DL $\approx$ 39900) — prawdopodobnie na skutek zanieczyszczenia próbek chloroformem

(1) W zestawieniach dotyczących wskaźników oznaczanych w laboratorium nie będą uwzględniane wartości uzyskane na drodze obliczeniowej — oznaczenia twardości węglanowej i dwutlenku węgla agresywnego.

w procesie opróbowania, przechowywania i transportu (Witczak et al., 1994). Z kolei w przypadku sodu, chlorków lub siarczanów stosunek PDL/DL = 1, co oznacza, że wyniki te cechują się zadowalającą precyzją.

Powyższe wyniki potwierdzają konieczność prowadzenia kontroli wartości PDL, a w razie niezadowalających wyników, potrzebę wykrycia błędów grubych i ich usunięcia, tak by zapewnić właściwy poziom PDL.

Tabela 1.3. Granice oznaczalności DL i praktyczne granice oznaczalności PDL (wg Witczak et al., 1994) oraz maksymalne dopuszczalne stężenia MDP wybranych wskaźników w wodach pitnych wg polskich norm (Dz.U. Nr 82 z 2000 r., poz. 937) i wytycznych Unii Europejskiej (Dyrektywa Unii Europejskiej 98/83/EC)

Lp.	Analizowana zmienna	Jednostka	DL		PDL		MDP
			Tarnów	Katowice	Tarnów	Katowice	
1.	Suma substancji rozpuszczonych	mg/dm^3	1	5	4	5	—
2.	Zasadowość ogólna	$mval/dm^3$	0.05	0.1	0.15	0.1	—
3.	Twardość ogólna	$mval/dm^3$	2	4	2	4	60–500
4.	Potas	mg/dm^3	0.01	0.2	1.5	0.7	—
5.	Sód	mg/dm^3	0.1	0.2	0.1	0.8	200
6.	Magnez	mg/dm^3	0.1	5	0.7	5	50
7.	Wapń	mg/dm^3	0.1	0.5	4.8	5	—
8.	Azot amonowy	mg/dm^3	0.04	0.01	0.1	0.15	0.5
9.	Glin	mg/dm^3	0.015	0.01	0.08	0.15	0.2
10.	Żelazo ogólne	mg/dm^3	0.03	0.01	2.2	0.27	0.2
11.	Mangan	mg/dm^3	0.01	0.01	0.03	0.085	0.05
12.	Azot azotynowy	mg/dm^3	0.001	0.001	0.003	0.007	0.1
13.	Azot azotanowy	mg/dm^3	0.1	0.1	0.2	—	50
14.	Chlorki	mg/dm^3	5	0.5	5	5	250
15.	Siarczany	mg/dm^3	10	10	10	10	250
16.	Fosforany rozpuszczone	mg/dm^3	0.05	0.05	0.05	0.05	5*
17.	Krzemionka zdysocjowana	mg/dm^3	0.7	0.5	0.7	0.5	—
18.	Fluorki	mg/dm^3	0.1	0.01	0.1	0.01	1.5
19.	Bor	mg/dm^3	0.005	—	0.23	—	1
20.	Chrom ogólny	mg/dm^3	0.003	0.01	0.01	0.01	0.05
21.	Cynk	mg/dm^3	0.01	0.01	0.075	0.11	3
22.	Kadm	mg/dm^3	0.001	0.001	0.001	0.001	0.003
23.	Miedź	mg/dm^3	0.002	0.01	0.01	0.01	1
24.	Nikiel	mg/dm^3	0.001	0.01	0.02	0.01	0.02
25.	Ołów	mg/dm^3	0.001	0.005	0.0068	0.016	0.01
26.	Rtęć	mg/dm^3	0.0002	0.0002	0.0037	0.0002	0.001
27.	Współczynnik absorpcji UV (A254)		0.005	0.005	0.032	—	—
28.	Rozpuszczony węgiel organiczny	mg/dm^3	0.2	0.5	2.33	—	—
29.	Utlenialność ChZT-Mn	mg/dm^3	0.5	0.5	1.12	—	5
30.	Azot organiczny Kjeldahla	mg/dm^3	0.5	0.5	0.5	0.5	—
31.	Fenole lotne	mg/dm^3	0.001	0.01	0.001	0.01	0.0005
32.	Substancje ropopochodne	mg/dm^3	0.05	0.01	0.45	—	—
33.	Chloroform	mg/dm^3	0.00001	0.0001	0.399	—	0.03
34.	Subst. pow.-czynne anionowe	mg/dm^3	0.0001	0.0001	0.0001	0.0001	0.2
35.	Czterochloroetylen	mg/dm^3	0.000005	0.000005	0.0056	—	0.01
36.	Trójchloroetylen	mg/dm^3	0.00003	0.00001	0.019	0.0028	0.05
37.	DDT	mg/dm^3	0.00002	0.000005	0.002	0.000005	—
38.	DDE	mg/dm^3	0.000008	0.000005	0.0004	0.000005	—
39.	DDD	mg/dm^3	0.000008	0.000005	0.003	0.000006	—
40.	Gamma-HCH	mg/dm^3	0.000008	0.000003	0.000009	0.000004	—
41.	Metoksychlor	mg/dm^3	0.00005	0.000008	0.001	0.000008	—
42.	Benzo-a-piren	mg/dm^3	—	—	—	—	—
43.	Suma 6WWA	mg/dm^3	—	—	—	—	—

*fosfor jako P_2O_5; znak — oznacza brak danych

1.1.2.2. Ocena precyzji oznaczeń

Precyzja jest jednym z najważniejszych parametrów określających jakość pomiarów analitycznych. Określa ona rozrzut wyników wokół centralnej wartości zbioru, którą stanowi średnia arytmetyczna z wyników pomiarów.

Termin zbiór w tym przypadku oznacza liczbę niezależnych powtórzeń pomiarów dokonywanych w warunkach powtarzalności lub odtwarzalności.

Zgodnie z Międzynarodowym słownikiem podstawowych i ogólnych terminów metrologii (1996):

- **Powtarzalność** (*repeatability*) — to stopień zgodności wyników pomiarów tej samej wielkości mierzonej (produktu, próbki), wykonywanych w tych samych warunkach pomiarowych. Obejmują one tę samą procedurę pomiarową, tego samego operatora, ten sam przyrząd pomiarowy stosowany w tych samych warunkach, to samo miejsce oraz krótkie odstępy czasu.
- **Odtwarzalność** (*reproducibility*) to stopień zgodności wyników pomiarów tej samej wielkości mierzonej (produktu, próbki), wykonywanych w zmienionych warunkach pomiarowych. Warunki podlegające zmianom mogą obejmować: zasadę pomiaru, metodę pomiaru, obserwatora, przyrząd pomiarowy, etalon odniesienia, miejsce, warunki stosowania oraz czas.

Do oceny precyzji używa się tych wszystkich charakterystyk, które w statystyce matematycznej skonstruowano do opisu rozproszenia wartości zmiennych losowych. Podstawową rolę odgrywa tu rozstęp danych R, wariancja s^2 i odchylenie standardowe s, a spośród względnych miar rozproszenia — współczynnik zmienności V. Odpowiednie wzory dotyczące wymienionych statystyk można znaleźć w dodatku B niniejszej pracy lub w literaturze (Doerffel, 1993; Helsel, Hirsch, 1992; Szczepańska, Kmiecik, 1998).

Ocena precyzji na podstawie wyników oznaczeń próbek dublowanych z wykorzystaniem analizy wariancji ANOVA

W monitoringu jakości wód podziemnych do oceny precyzji wyników stosuje się analizę wariancji ANOVA, gdyż w przeciwieństwie do odchylenia standardowego (stosowanego zazwyczaj do oceny precyzji) wariancja jest addytywna i daje się sumować, pod warunkiem, że źródła wariancji są niezależne (Ramsey, 1992; Ramsey et al., 1992).

Dotychczasowe doświadczenia związane z realizacją terenowego programu kontroli QA/QC (Witczak et al., 1994, 1994a; Osmęda-Ernst et al., 1995, 1996; Szczepańska et al., 1996, 1996a, 1997) wskazują, że jest to najtańsza, najszybsza i najlepsza metoda oszacowania błędów losowych powstających w procesach opróbowania i/lub analityki.

Naturalna zmienność hydrogeochemiczna wyrażona za pomocą wariancji hydrogeochemicznej (σ_g^2) zaburzana jest przez dwa procesy, w których powstają błędy losowe: opróbowanie dające dodatkową zmienność w postaci wariancji opróbowania (σ_s^2) oraz analizę chemiczną z dodatkową wariancją analityczną (σ_a^2).

Wobec tego, zgodnie ze wzorem podanym przez Ramseya et al. (1992), całkowitą obserwowaną zmienność przestrzenną można przedstawić w formie wariancji całkowitej (σ_{tot}^2) będącej sumą wymienionych powyżej wariancji:

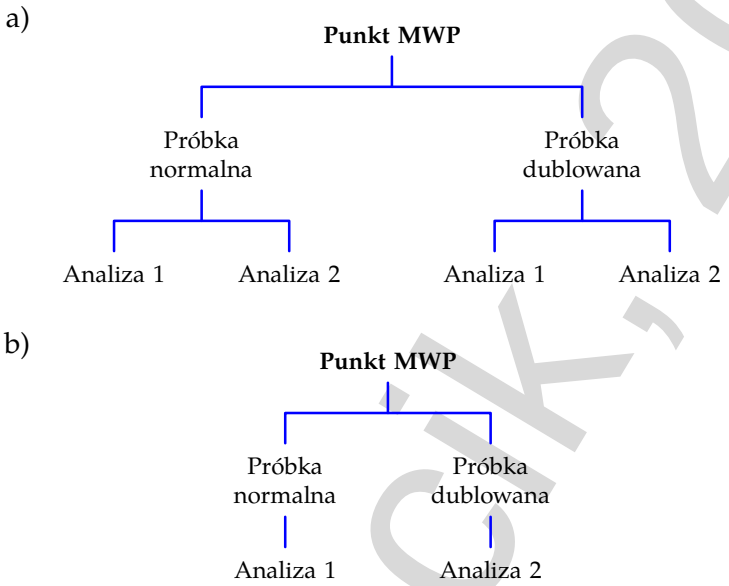
$$\sigma_{tot}^2 = \sigma_g^2 + \sigma_s^2 + \sigma_a^2 \tag{1.7}$$

Rozdzielenie błędów losowych powstałych w trakcie opróbowania (σ_s^2) od błędów analitycznych (σ_a^2) wymaga realizacji rozszerzonego laboratoryjnego programu badań próbek normalnych i dublowanych (rys. 1.8a). W przypadku gdy realizacja rozszerzonego programu przedstawionego na rysunku 1.8a jest niemożliwa (najczęściej ze względów finansowych), prowadzi się opróbowanie punktu monitoringowego i badania analityczne zgodnie ze schematem podanym na rysunku 1.8b.

Ocenia się wówczas tak zwaną wariancję techniczną σ^2_{tech}, która uwzględnia łączny błąd opróbowania (σ^2_s) i analityki (σ^2_a):

$$\sigma^2_{tech} = \sigma^2_s + \sigma^2_a \tag{1.8}$$

Precyzja oznaczeń jest zadowalająca, jeśli udział wariancji technicznej (σ^2_{tech}) w całkowitej wariancji (σ^2_{tot}) nie przekracza maksymalnego dopuszczalnego poziomu wynoszącego 20%.



Rysunek 1.8. Schemat opróbowania punktu monitoringowego i badań analitycznych stosowanych do oceny precyzji wyników za pomocą analizy wariancji (wg Ramseya, 1992): a) schemat badań umożliwiający ocenę wariancji opróbowania σ^2_s i wariancji analitycznej σ^2_a; b) schemat badań pozwalający ocenić tzw. wariancję techniczną σ^2_{tech}

Terenowy program kontroli jakości QA/QC przeprowadzony w I serii opróbowania sieci RMWP dorzecza górnej Wisły, zgodnie ze schematem podanym na rysunku 1.8b, umożliwił obliczenie za pomocą programu komputerowego ROB 2 (tab. 1.4), wariancji technicznej σ^2_{tech}, uwzględniającej łączny wpływ błędów opróbowania i analityki.

Tabela 1.4. Wyliczone za pomocą programu ROB 2 odchylenia standardowe dla cynku, przy zastosowaniu procedury klasycznej ANOVA i robust statistics (opróbowanie 1993 r., I seria badań)

```
--- element CYNK
Classical results: Mean = 1.0420833E-01
Sums of Squares are - 0.9548939 0.7574359 0.0000000E+00
sigma values(geochem, sampling, analysis) - 0.071 0.178 0.000
sigma (total) - 0.191
Robust results:
mean = 7.2811484E-02
sigma values(geochem, sampling, analysis) - 0.058 0.019 0.000
sigma (total) - 0.061
```

Program ROB 2 pozwala na oszacowanie wariancji metodą klasyczną i metodą *robust statistics* (elastycznego postępowania statystycznego, bez odrzucania błędów grubych). Aby informacje uzyskane za pomocą programu były dostatecznie wiarygodne, obliczenia powinny być prowadzone dla co najmniej 11 par próbek (próbka normalna i dublowana).

W przypadku wyników pomiarów niższych od granicy oznaczalności DL do obliczeń wykorzystywano wartości liczbowe DL ($< \mathrm{DL} = \mathrm{DL}$). W obliczeniach nie uwzględniono tych

par, dla których wyniki oznaczeń w obu próbkach: normalnej i dublowanej, były niższe od granicy oznaczalności (ich uwzględnienie spowodowałoby nieuzasadniony wzrost precyzji wyników badań hydrogeochemicznych — obniżenie σ^2_{tech}).

W tabelach 1.5 i 1.6 zestawiono procentowe udziały wariancji technicznej w wariancji całkowitej, obliczone metodą klasycznej analizy wariancji ANOVA oraz elastycznego postępowania statystycznego ROBUST.

Tabela 1.5. RMWP dorzecza górnej Wisły — I seria opróbowania. Wariancja techniczna obliczona metodą klasycznej analizy wariancji jako procent wariancji całkowitej. Brak wartości σ^2_{tech} oznacza brak danych ze względu na $N < 11$ liczbę par wyników dla próbek normalnych i dublowanych

Lp.	Analizowana zmienna	Jednostka	N	σ^2_{tech} ANOVA [%]
1.	Siarczany	mg/dm^3	24	2.04
2.	Fosforany rozpuszczone	mg/dm^3	17	2.10
3.	Azot azotynowy	mg/dm^3	22	2.78
4.	Zasadowość ogólna	$mval/dm^3$	24	3.47
5.	Fluorki	mg/dm^3	24	3.54
6.	Twardość ogólna	mg $CaCO_3/dm^3$	24	3.61
7.	Wapń	mg/dm^3	24	4.09
8.	Krzemionka zdysocjowana	mg/dm^3	24	4.45
9.	Substancje ropopochodne	mg/dm^3	24	4.57
10.	Utlenialność ChZT-Mn	mg/dm^3	24	5.35
11.	Suma substancji rozpuszczonych	mg/dm^3	24	5.36
12.	Azot azotanowy	mg/dm^3	19	7.17
13.	Azot amonowy	mg/dm^3	13	7.99
14.	Współczynnik absorpcji UV (A254)		24	8.37
15.	Nikiel	mg/dm^3	12	11.11
16.	Gamma-HCH	mg/dm^3	17	11.11
17.	Rozpuszczony węgiel organiczny	mg/dm^3	23	12.86
18.	Chlorki	mg/dm^3	23	15.36
19.	Sód	mg/dm^3	24	15.39
20.	DDD	mg/dm^3	20	17.73
21.	Magnez	mg/dm^3	24	18.04
22.	Chloroform	mg/dm^3	24	22.14
23.	Miedź	mg/dm^3	19	25.00
24.	DDE	mg/dm^3	20	25.00
25.	Rtęć	mg/dm^3	14	27.39
26.	Glin	mg/dm^3	23	27.56
27.	Metoksychlor	mg/dm^3	15	29.34
28.	DDT	mg/dm^3	15	32.65
29.	Żelazo ogólne	mg/dm^3	24	45.29
30.	Potas	mg/dm^3	24	50.02
31.	Azot organiczny Kjeldahla	mg/dm^3	22	58.69
32.	Fenole lotne	mg/dm^3	11	64.00
33.	Mangan	mg/dm^3	19	68.56
34.	Ołów	mg/dm^3	22	69.44
35.	Cynk	mg/dm^3	24	86.85
36.	Czterochloroetylen	mg/dm^3	22	100.00
37.	Trójchloroetylen	mg/dm^3	19	100.00
38.	Bor	mg/dm^3	10	
39.	Chrom ogólny	mg/dm^3	6	
40.	Kadm	mg/dm^3	7	
41.	Subst. pow.-czynne anionowe	mg/dm^3	9	
42.	Benzo-a-piren	mg/dm^3	10	
43.	Suma 6WWA	mg/dm^3	0	

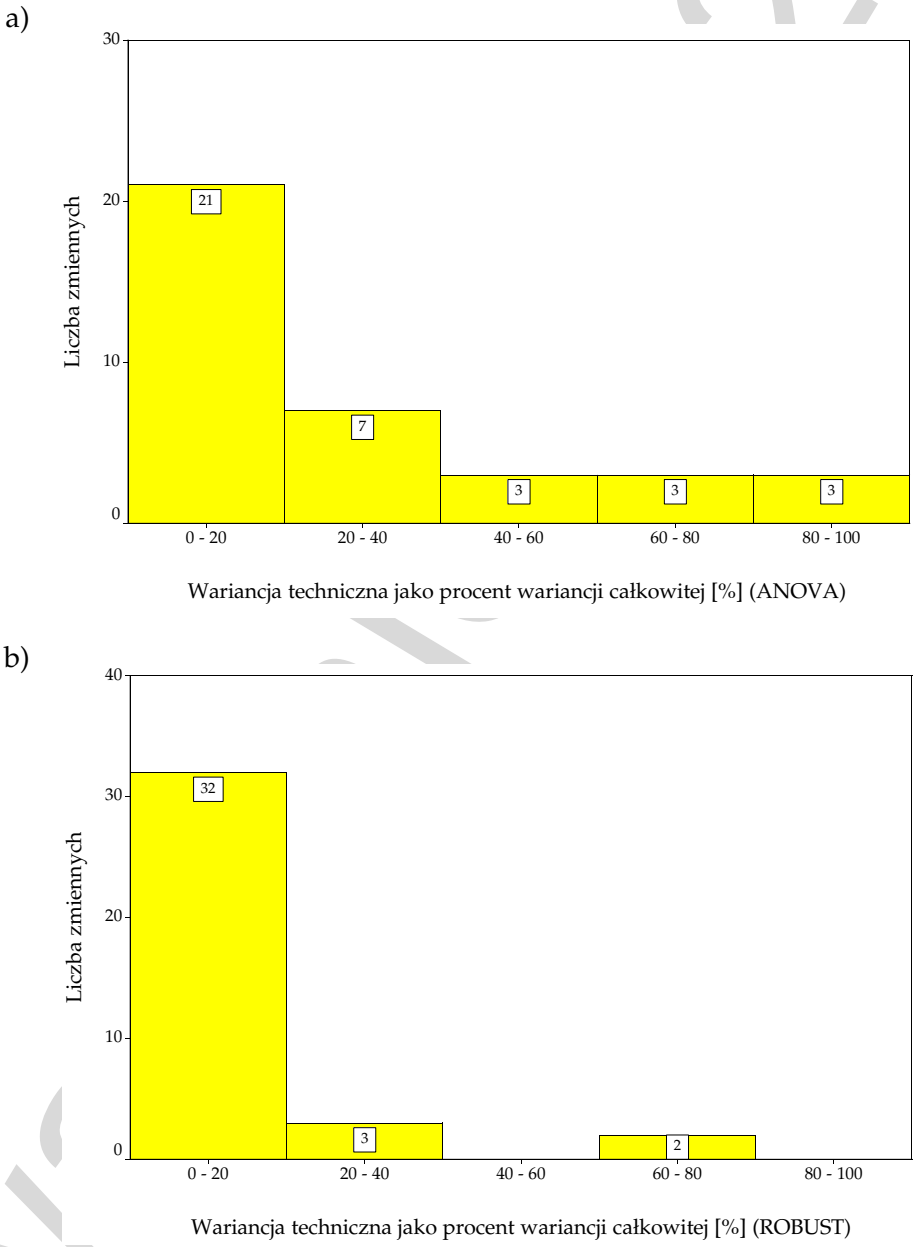
W wyniku elastycznego postępowania statystycznego w większości przypadków uzyskuje się niższą wariancję techniczną, co oznacza, że wyniki pomiarów badanych wskaźników obarczone są błędami grubymi. Wyższą wariancję zanotowano jedynie w przypadku azotu amonowego, rozpuszczonego węgla organicznego, azotu azotynowego i niklu.

Tabela 1.6. RMWP dorzecza górnej Wisły — I seria opróbowania. Wariancja techniczna obliczona metodą statystyk robust jako procent wariancji całkowitej. Brak wartości σ^2_{tech} oznacza brak danych ze względu na $N < 11$ liczbę par wyników dla próbek normalnych i dublowanych

Lp.	Analizowana zmienna	Jednostka	N	σ^2_{tech} ROBUST [%]
1.	DDT	mg/dm^3	15	0.00
2.	DDE	mg/dm^3	20	0.00
3.	Gamma-HCH	mg/dm^3	17	0.00
4.	Metoksychlor	mg/dm^3	15	0.00
5.	Zasadowość ogólna	$mval/dm^3$	24	0.15
6.	Mangan	mg/dm^3	19	0.20
7.	Fluorki	mg/dm^3	24	0.21
8.	Suma substancji rozpuszczonych	mg/dm^3	24	0.28
9.	Krzemionka zdysocjowana	mg/dm^3	24	0.43
10.	Twardość ogólna	mg $CaCO_3/dm^3$	24	0.45
11.	Siarczany	mg/dm^3	24	0.57
12.	Azot azotanowy	mg/dm^3	19	0.61
13.	Wapń	mg/dm^3	24	0.75
14.	Chlorki	mg/dm^3	23	1.19
15.	Sód	mg/dm^3	24	1.29
16.	Fosforany rozpuszczone	mg/dm^3	17	1.77
17.	Żelazo ogólne	mg/dm^3	24	1.84
18.	Substancje ropopochodne	mg/dm^3	24	2.30
19.	Współczynnik absorpcji UV (A254)		24	2.37
20.	Rtęć	mg/dm^3	14	3.24
21.	Utlenialność ChZT-Mn	mg/dm^3	24	5.04
22.	Potas	mg/dm^3	24	5.82
23.	Magnez	mg/dm^3	24	5.91
24.	Chloroform	mg/dm^3	24	6.16
25.	Azot organiczny Kjeldahla	mg/dm^3	22	6.17
26.	Glin	mg/dm^3	23	6.25
27.	Miedź	mg/dm^3	19	6.25
28.	Trójchloroetylen	mg/dm^3	19	6.25
29.	Cynk	mg/dm^3	24	9.70
30.	DDD	mg/dm^3	20	11.11
31.	Azot amonowy	mg/dm^3	13	13.22
32.	Rozpuszczony węgiel organiczny	mg/dm^3	23	18.28
33.	Azot azotynowy	mg/dm^3	22	25.00
34.	Nikiel	mg/dm^3	12	25.00
35.	Ołów	mg/dm^3	22	36.00
36.	Fenole lotne	mg/dm^3	11	64.00
37.	Czterochloroetylen	mg/dm^3	22	69.45
38.	Bor	mg/dm^3	10	
39.	Chrom ogólny	mg/dm^3	6	
40.	Kadm	mg/dm^3	7	
41.	Subst. pow.-czynne anionowe	mg/dm^3	9	
42.	Benzo-a-piren	mg/dm^3	10	
43.	Suma 6WWA	mg/dm^3	0	

Dla sześciu wskaźników spośród czterdziestu trzech oznaczanych w laboratorium (bor, chrom ogólny, kadm, substancje powierzchniowo-czynne, benzo-a-piren, suma 6WWA) nie można było wyznaczyć poziomu wariancji technicznej ze względu na niedostateczną ($N < 11$)

liczbę par próbek normalnych i dublowanych. W dwudziestu jeden przypadkach (tab. 1.6) poziom wariancji technicznej wyznaczonej z wykorzystaniem klasycznej analizy wariancji ANOVA jest zadowalający ($\sigma^2_{tech} < 20\%$). W siedmiu przypadkach (glin, miedź, rtęć, chloroform, DDT, DDE, metoksychlor) wariancja techniczna kształtowała się na poziomie ok. 30% wariancji całkowitej. W przypadku potasu, żelaza ogólnego i azotu organicznego uzyskano wyniki w przedziale 40–60% zmienności całkowitej. Wariancją techniczną powyżej 60% wariancji całkowitej charakteryzowały się wyniki oznaczeń manganu, ołowiu i fenoli lotnych. Najniższą precyzją — wariancja techniczna stanowi ponad 80% wariancji całkowitej — charakteryzowały się wyniki oznaczeń cynku, czterochloroetylenu i trójchloroetylenu (rys. 1.9a).



Rysunek 1.9. Rozkład wariancji technicznej σ^2_{tech} obliczonej za pomocą klasycznej analizy wariancji (a) i z wykorzystaniem statystyk robust (b)

Po zastosowaniu elastycznego postępowania statystycznego uzyskano niższe udziały procentowe wariancji technicznej w wariancji całkowitej (rys. 1.9b). W trzydziestu dwóch przypadkach poziom wariancji technicznej nie przekraczał 20% wariancji całkowitej. Oznaczenia

azotu azotynowego, niklu i ołowiu charakteryzowała wariancja techniczna w granicach od 20 do 40%. Jeszcze większą zmiennością charakteryzowały się wyniki oznaczeń fenoli i czterochloroetylenu — wariancja techniczna stanowi w tym przypadku 60–70% wariancji całkowitej.

Do prognozowania zmian jakości wód w układzie przestrzennym mogą być wykorzystane wiarygodne wyniki oznaczeń wskaźników fizyko-chemicznych. Wskaźniki chemiczne wód, dla których wariancja techniczna przekracza dopuszczalny poziom 20% należy wyłączyć ze zbioru, na którym oparte będzie prognozowanie zmian jakości wód, gdyż błędy w bazie danych wejściowych skutkują powielaniem ich w prognozach dotyczących zmian jakości wód.

Ocena precyzji w warunkach powtarzalności i odtwarzalności

W przypadku oznaczeń mikroskładników, których stężenia w wodach są z reguły niskie i bardzo często zbliżone do granicy oznaczalności DL, konieczne jest prowadzenie bieżącej kontroli jakości badań za pomocą kart kontrolnych lub innych testów ułatwiających wyodrębnienie obserwacji obarczonych błędami grubymi (Szczepańska, Kmiecik, 1998).

Wyniki pomiarów, dla których obserwujemy sygnały na kartach kontrolnych nie mogą być uwzględniane w obliczeniach zarówno precyzji oznaczeń, jak też i innych parametrów kontroli jakości (np. DL, PDL).

Dane literaturowe wskazują, że o jakości uzyskanych wyników pomiarów badanych wskaźników fizyko-chemicznych wód decydują głównie: proces opróbowania (Nielsen, 1991), precyzja zastosowanej metody analitycznej i warunki, w jakich wykonywane są oznaczenia (powtarzalność i odtwarzalność pomiarów). To skłoniło autorkę do podjęcia dodatkowych badań, których obiektem były podziemne wody jurajskie ze Zdroju Królewskiego w Krakowie. Do szczegółowych rozważań wybrano wyniki oznaczeń cynku (przedstawiciela metali ciężkich), który z uwagi na łatwość migracji powszechnie występuje w wodach podziemnych (Macioszczyk, 1987).

Przeprowadzono analizę precyzji oznaczeń (w warunkach powtarzalności i odtwarzalności) tego wskaźnika w próbkach ze Zdroju Królewskiego w Krakowie (Kleczkowski et al., 1994; Szczepańska, Kmiecik, 1998a, 1998b).

Próbki z tego zdroju były pobierane do badań w ramach programu kontroli jakości QA/QC regionalnego monitoringu jakości wód podziemnych dorzecza górnej Wisły w 1993 roku (Witczak et al., 1994), w ramach okresowej kontroli jakości tych wód przeprowadzonej w 1996 roku (Osmęda-Ernst et al., 1996) oraz w ramach systematycznych badań jakości wód jurajskich ze Zdroju Królewskiego w latach 1998–2000. Łącznie pobrano 66 próbek, które posłużyły do oceny precyzji oznaczeń cynku w wodach podziemnych.

W roku 1993, w ramach regionalnego monitoringu jakości wód podziemnych dorzecza górnej Wisły pobrano 20 próbek jurajskiej wody podziemnej ze Zdroju Królewskiego w Krakowie (tab. 1.7). Opróbowanie prowadzono za pomocą sprzętu wielokrotnego użytku firmy Eijkelkamp, rutynowo stosowanego w monitoringu jakości wód podziemnych (Witczak, Adamczyk, 1994).

W ramach okresowej kontroli jakości wód pitnych z wapieni jury, w roku 1996 z tego samego Zdroju pobrano 9 próbek wód podziemnych. Proces opróbowania prowadzony był wówczas za pomocą sprzętu jednorazowego użytku firmy Millipore.

W latach 1998–2000 pobrano 37 próbek wód podziemnych ze zdroju Królewskiego. Opróbowanie prowadziło dwóch operatorów (15 próbek pobrano w roku 1998/1999 zaś 22 próbki w roku 1999/2000), również za pomocą sprzętu jednorazowego użytku firmy Millipore a badania składu chemicznego wykonywane były w laboratorium Zakładu Hydrogeologii i Ochrony Wód AGH.

Opróbowanie wód ze Zdroju Królewskiego we wszystkich przypadkach prowadzono zgodnie z zasadami przyjętymi w monitoringu jakości wód podziemnych (Witczak, Adamczyk, 1994; Prawo..., 1996), gdzie jako standard przyjmuje się filtrację próbki *on-line* w terenie, przez filtr membranowy $\phi = 0.45$ μm. Oznacza to, że w badaniach jakości wód podziemnych

analizowane są składniki rozpuszczone w wodzie. Próbki do badań były utrwalane w miejscu ich poboru, zgodnie z procedurami stosowanymi w monitoringu jakości wód podziemnych (Witczak, Adamczyk, 1994, 1995).

W laboratoriach WIOŚ Tarnów i OBiKŚ Katowice cynk był oznaczany przy zastosowaniu tej samej metody pomiarowej — absorpcyjnej spektrofotometrii atomowej AAS (zgodnie z PN-92/C-04570.02), obydwa laboratoria deklarowały identyczne granice oznaczalności DL = 0.01 mg/dm^3. W laboratorium ZHiOW AGH cynk oznaczano metodą atomowej spektrometrii emisyjnej z plazmą wzbudzoną indukcyjnie (ICP-AES), deklarowana przez laboratorium granica oznaczalności cynku w wodach DL = 0.005 mg/dm^3 (Osmęda-Ernst et al., 1995).

Zdrój Królewski jest zlokalizowany w Krakowie, na północ od Parku Krakowskiego, w południowej części skwerku należącego do Placu Inwalidów i na południe od ulicy Królewskiej, w odległości kilkudziesięciu metrów od niej. Otwór ujmujący wodę jurajską posiada 85.0 m głębokości, a jego profil przedstawia się następująco: utwory czwartorzędowe do głębokości 16.5 m. ppt., miocenu od 16.5 do 60.7 m. ppt. i utwory jurajskie od 60.7 do 85.0 m. ppt. Woda jurajska posiada stabilny skład chemiczny, nawet niefiltrowana próbka przechowywana w plastikowych naczyniach przez 5 dni nie zmienia swojego składu. Woda filtrowana (filtr membranowy 0.45 μm) ma skład chemiczny prawie taki sam jak niefiltrowana, zmniejsza się jedynie zawartość żelaza i cynku. Istotny jest fakt, iż woda ta, z uwagi na występowanie izolującej pokrywy w postaci iłów mioceńskich nie jest podatna na zanieczyszczenia antropogeniczne.

Stężenie cynku w wodzie ze Zdroju Królewskiego kształtuje się na względnie stałym poziomie 0.04 mg/dm^3 w próbkach niefiltrowanych, podczas gdy w wodzie filtrowanej jest niższe i wynosi 0.01 mg/dm^3 (Kleczkowski et al., 1994).

Tabela 1.7. Zestawienie ilościowe próbek analizowanych w laboratoriach WIOŚ Tarnów, OBiKŚ Katowice i ZHiOW AGH. Woda jurajska ze Zdroju Królewskiego w Krakowie. Objaśnienia: DL — deklarowana przez laboratorium granica oznaczalności cynku

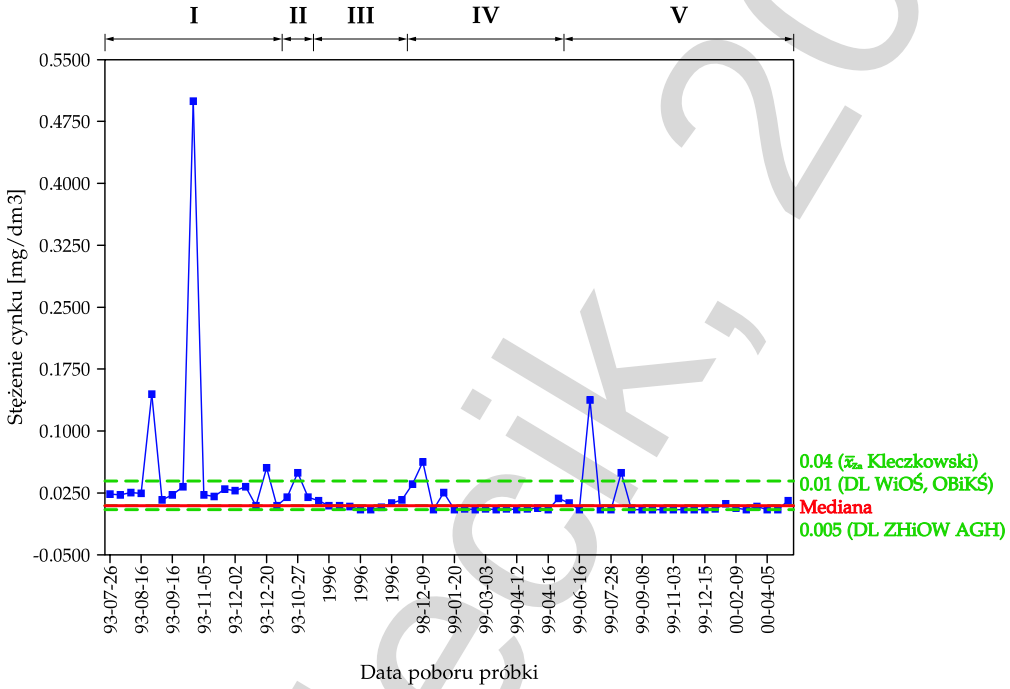
Rok opróbowania	Liczba próbek	Nr podzbioru	Laboratorium	Metoda oznaczania	DL
Źródło danych Witczak et al. (1994):					
1993	17	I	WIOŚ Tarnów	AAS	0.01 mg/dm^3
1993	3	II	OBiKŚ Katowice	AAS	0.01 mg/dm^3
				Suma:	20 próbek
Źródło danych Osmęda-Ernst et al. (1996):					
1996	9	II	WIOŚ Tarnów	AAS	0.01 mg/dm^3
				Suma:	9 próbek
Źródło danych Bieniek (1999):					
1998/1999	15	IV	ZHiOW AGH	ICP-AES	0.005 mg/dm^3
Źródło danych Pasiut (2000):					
1999/2000	22	V	ZHiOW AGH	ICP-AES	0.005 mg/dm^3
				Suma:	37 próbek
				RAZEM:	**66 próbek**

Analiza wyników oznaczeń cynku w próbkach wody jurajskiej ze Zdroju Królewskiego

Analizę wyników oznaczeń cynku w próbkach pobranych ze Zdroju Królewskiego w Krakowie prowadzono za pomocą programów SPSS PL v. 10.0 i QI Analyst 3.5 DB (SPSS, 1997, 1999). Program QI Analyst dostarcza narzędzi do szeroko pojętej kontroli jakości badań, procesów i produktów. Pozwala on na tworzenie kart przebiegu, histogramów, wykresów Pareto i kart kontrolnych różnego typu (karty kontrolne dla zmiennych, karty kontrolne dla atrybutów).

Program ma dobre możliwości prezentacji graficznej uzyskanych wyników, w łatwy sposób można go skonfigurować, przystosowując do własnych potrzeb informacje wyświetlane na kartach. Pozwala na pracę w środowisku sieciowym. Szczegóły dotyczące programu można znaleźć w dokumentacji (SPSS, 1997a).

Wyniki oznaczeń cynku we wszystkich próbkach ($N = 66$) jurajskiej wody pitnej ze Zdroju Królewskiego (tab. 1.7) przedstawiono na karcie przebiegu (rys. 1.10).



Rysunek 1.10. Wyniki oznaczeń cynku [mg/dm^3] w jurajskiej wodzie pitnej. Zdrój Królewski w Krakowie. Karta przebiegu. Metoda AAS: I — Tarnów 1993, II — Katowice 1993, III — Tarnów 1996; metoda ICP-AES: IV — ZHiOW AGH 1998/1999, V — ZHiOW AGH 1999/2000

Próbki analizowane w poszczególnych laboratoriach naniesione są na kartę przebiegu w układzie chronologicznym, zgodnie z czasem wykonywania analiz. Ma to umożliwić śledzenie ewentualnych zmian jakości badań w funkcji czasu. W omawianym przypadku stężenia dla kilku próbek wyraźnie odbiegają od pozostałych.

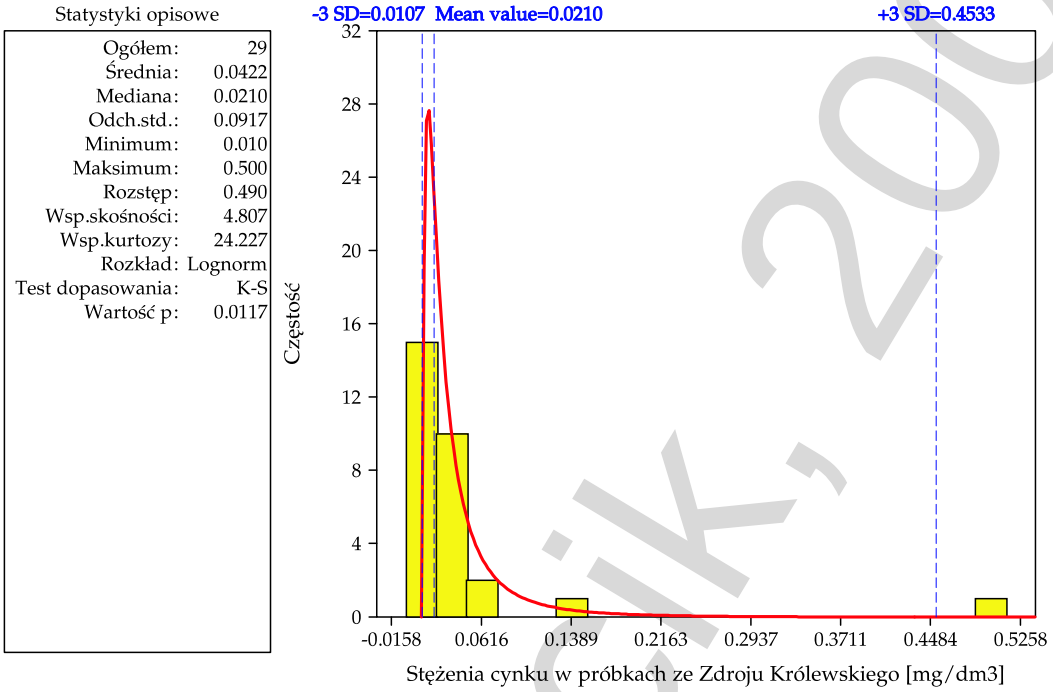
Ze względu na różnice w metodyce wykonywanych w obu laboratoriach analiz (różne deklarowane przez laboratoria granice oznaczalności $DL_{AAS} = 0.01 mg/dm^3$, $DL_{ICP\text{-}AES} = 0.005$ mg/dm^3 — tab. 1.7) do dalszej analizy wyniki podzielono na dwie grupy:

- w skład pierwszej z grup wchodzą wyniki uzyskane w laboratorium WIOŚ Tarnów w roku 1993 (I podzbiór), wyniki z laboratorium OBiKŚ Katowice (II podzbiór) i wyniki uzyskane w laboratorium w Tarnowie w 1996 roku (III podzbiór);
- drugą grupę stanowią wyniki uzyskane w laboratorium ZHiOW AGH (IV i V podzbiór, lata 1998/1999 oraz 1999/2000).

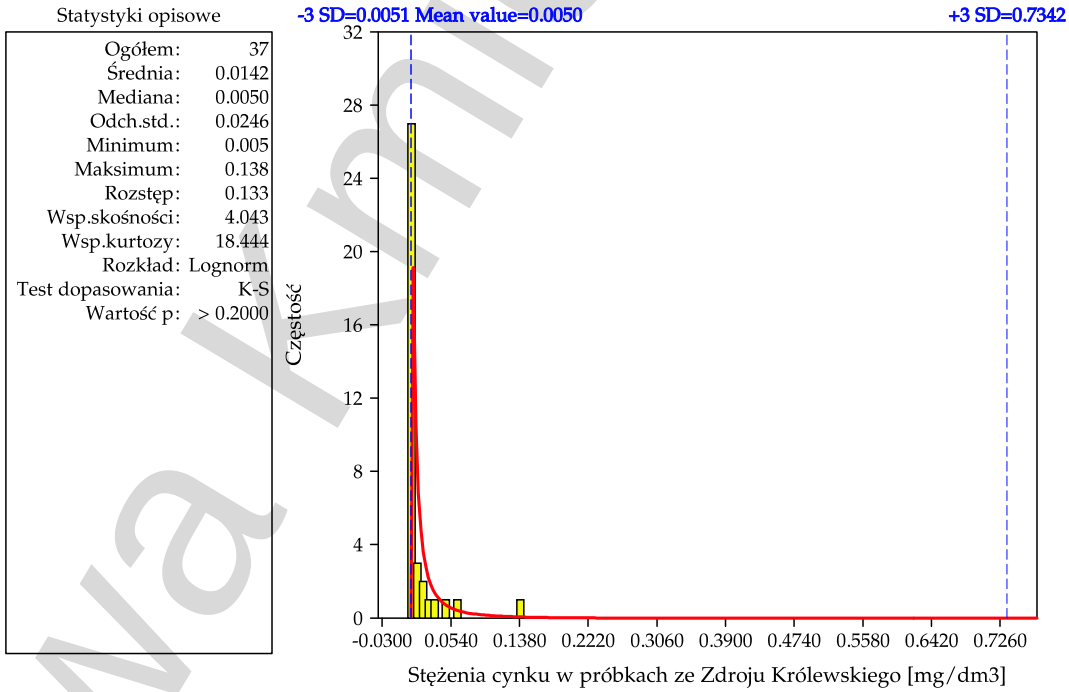
Na kartach kontrolnych pojedynczych pomiarów (rys. 1.13) obserwujemy trzy sygnały punktowe dla próbek nr 9 (Tarnów 1993), 31 (ZHiOW AGH 1998/1999) i 47 (ZHiOW AGH 1999/2000). Łączy się to przypuszczalnie z niewłaściwie prowadzoną filtracją próbki w terenie, w wyniku czego zawiesina zawierająca cynk przedostaje się do pojemnika z wodą.

Również Kleczkowski i inni (1994) podkreślają, iż proces filtracji wpływa na zmianę stężeń cynku w próbkach wody jurajskiej ze Zdroju Królewskiego. W próbkach poddanych procesowi filtracji obserwowano obniżanie stężeń cynku (0.01 mg/dm^3, 0.005 mg/dm^3) w stosunku do wartości notowanych dla próbek niefiltrowanych (0.04 mg/dm^3).

Wyniki oznaczeń cynku w obu grupach mają rozkład logarytmiczno-normalny (rys. 1.11, rys. 1.12), widoczne są wyniki poniżej granic oznaczalności DL oraz obserwacje odstające.

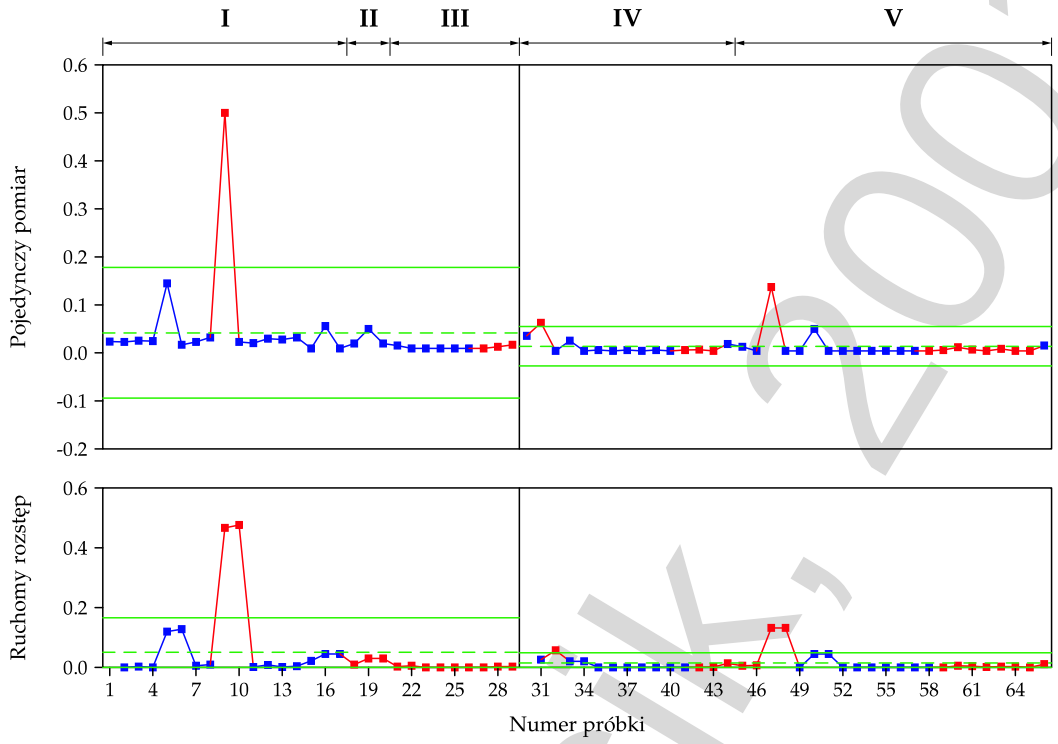


Rysunek 1.11. Wyniki oznaczeń cynku [mg/dm^3] w jurajskiej wodzie pitnej metodą AAS. Zdrój Królewski w Krakowie. Histogram rozkładu pojedynczych pomiarów wyników uzyskanych w laboratoriach WIOŚ Tarnów (1993, 1996) i OBiKŚ Katowice (1993)

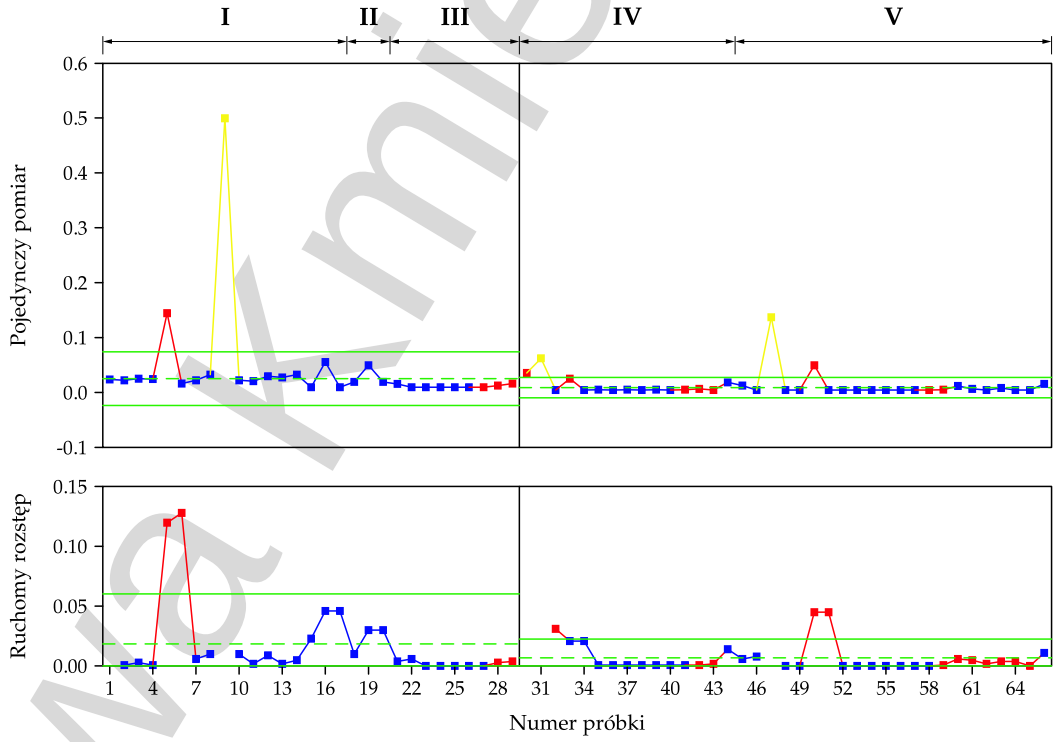


Rysunek 1.12. Wyniki oznaczeń cynku [mg/dm^3] w jurajskiej wodzie pitnej metodą ICP-AES. Zdrój Królewski w Krakowie. Histogram rozkładu pojedynczych pomiarów wyników uzyskanych w laboratorium ZHiOW AGH (1998–2000)

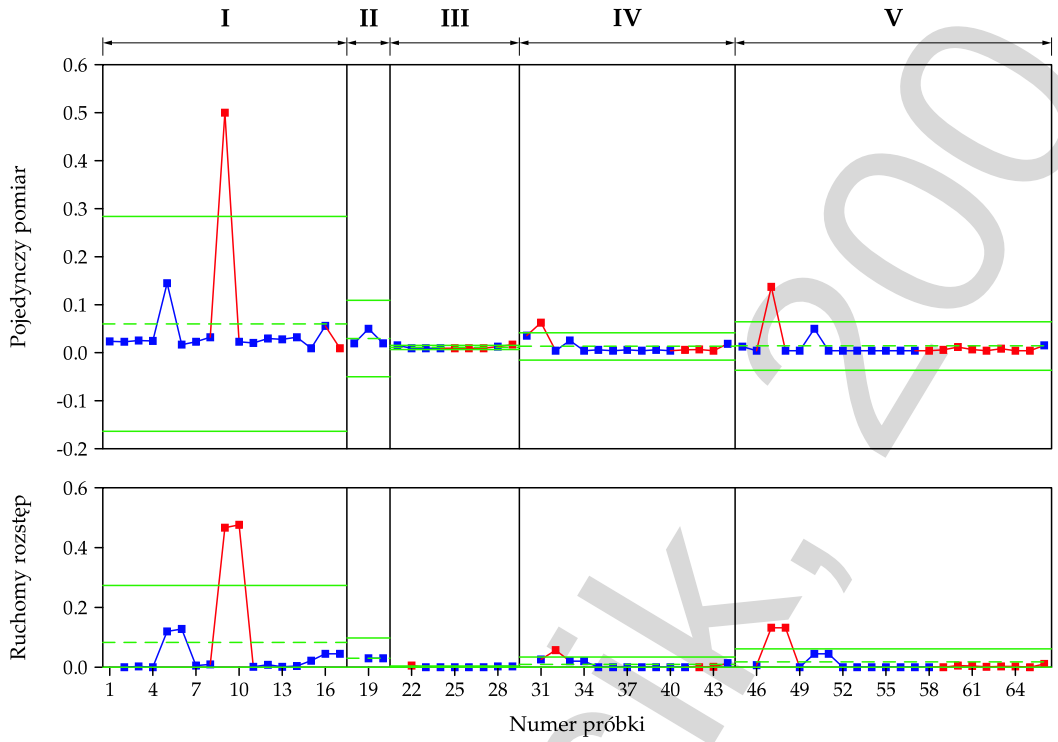
Próbka wody po filtracji poddawana jest dalej utrwaleniu (zakwaszeniu 5 ml stężonego HNO_3 na 1000 ml wody), zgodnie z procedurą stosowaną w monitoringu jakości wód pod-



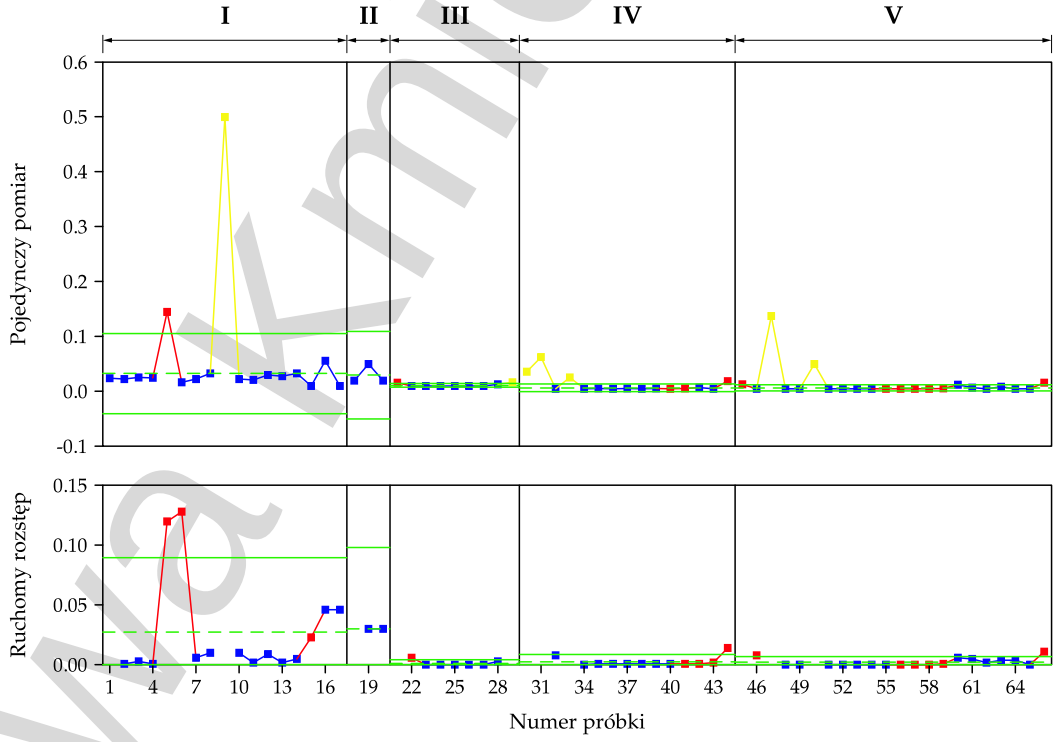
Rysunek 1.13. Wyniki oznaczeń cynku [mg/dm^3] w jurajskiej wodzie pitnej. Zdrój Królewski w Krakowie. Karta kontrolna pojedynczych pomiarów i ruchomych rozstępów — czerwonym kolorem zaznaczone są sygnały punktowe dla próbek 9, 31 i 47. Metoda AAS: I — Tarnów 1993, II — Katowice 1993, III — Tarnów 1996; metoda ICP-AES: IV — ZHiOW AGH 1998/1999, V — ZHiOW AGH 1999/2000



Rysunek 1.14. Wyniki oznaczeń cynku [mg/dm^3] w jurajskiej wodzie pitnej. Zdrój Królewski w Krakowie. Karta kontrolna pojedynczych pomiarów i ruchomych rozstępów po wyłączeniu sygnałów punktowych (próbki nr 9, 31, 47, zaznaczone na wykresie kolorem żółtym). Metoda AAS: I — Tarnów 1993, II — Katowice 1993, III — Tarnów 1996; metoda ICP-AES: IV — ZHiOW AGH 1998/1999, V — ZHiOW AGH 1999/2000



Rysunek 1.15. Wyniki oznaczeń cynku [mg/dm^3] w jurajskiej wodzie pitnej. Zdrój Królewski w Krakowie. Karty kontrolne pojedynczych pomiarów i ruchomych rozstępów dla analizowanych podzbiorów danych. Metoda AAS: I — Tarnów 1993, II — Katowice 1993, III — Tarnów 1996; metoda ICP-AES: IV — ZHiOW AGH 1998/1999, V — ZHiOW AGH 1999/2000



Rysunek 1.16. Wyniki oznaczeń cynku [mg/dm^3] w jurajskiej wodzie pitnej. Zdrój Królewski w Krakowie. Karty kontrolne pojedynczych pomiarów i ruchomych rozstępów dla analizowanych podzbiorów danych po wyłączeniu z analizy próbek nr 9, 29, 30, 31, 33, 47 i 50. Oznaczenia jak na rysunku powyżej

ziemnych (Witczak, Adamczyk, 1994, 1995), co powoduje, że przy niewłaściwie prowadzonej filtracji następuje przejście cynku z zawiesiny do roztworu — znajduje to odzwierciedlenie w podwyższonych stężeniach tego wskaźnika w próbkach analizowanej wody (sygnały punktowe na kartach kontrolnych).

Nie bez znaczenia jest również, zwłaszcza w przypadku mikroskładników, rodzaj zastosowanej metodyki oznaczeń. W omawianym przypadku wyraźnie widać, że dla metody analitycznej o niższej granicy oznaczalności (ICP-AES, DL = 0.005 mg/dm^3, podzbiory IV i V) uzyskano mniejszy rozrzut wyników pomiarów.

Wyniki tych badań wskazują jednoznacznie na znaczący udział procesu opróbowania w generowaniu błędów przypadkowych — grubych, rzutujących z kolei na precyzję oznaczeń analizowanego składnika (cynku).

Na kartach kontrolnych (rys. 1.13) widoczne są ponadto sygnały sekwencyjne, kilka kolejnych punktów leży poniżej linii centralnej np. próbki: 1–4, 21–29, 34–43, 52–65.

Po wyłączeniu z analizy próbek nr 9, 31 i 47 (sygnałów punktowych) karty mają postać jak na rysunku 1.14. Obserwujemy zmniejszenie zakresu granic kontrolnych — a więc poprawę precyzji oznaczeń.

Po podzieleniu całego zbioru danych na pięć podzbiorów — pod względem miejsca (WIOŚ Tarnów, OBiKŚ Katowice, ZHiOW AGH) i czasu wykonania analiz (1993, 1996, 1998/1999, 1999/2000) karty kontrolne dla poszczególnych podzbiorów wartości przedstawiają się jak na rysunku 1.15. Najmniejszy rozrzut cechuje wyniki w podzbiorach III–V.

Nadal widoczne są sygnały punktowe dla próbek nr 9, 31 i 47. Pojawia się także sygnał dla próbki nr 29. W podzbiorach I, IV i V obserwuje się pewną niestabilność kart w kilku początkowych pomiarach.

Po przejrzeniu dokumentacji dotyczącej poboru próbek wyłączono z analizy dodatkowe próbki: 30 i 33 w podzbiorze IV, oraz 50 w podzbiorze V (rys. 1.15).

Po ich wyłączeniu z analizy (rys. 1.16) precyzja oznaczeń w wydzielonych podzbiorach znacznie się poprawia, przy czym w podzbiorach III–V rozrzut jest minimalny.

Ocena precyzji oznaczeń cynku w próbkach wody jurajskiej ze Zdroju Królewskiego, w warunkach odtwarzalności

Badania próbek ze Zdroju Królewskiego wykonywały trzy laboratoria, każda z próbek była analizowana tylko raz, nie można więc przeprowadzić pełnej analizy powtarzalności i odtwarzalności systemu pomiarowego za pomocą np. programu Gage R&R czy Gage Manager (SPSS, 1997a; Kmiecik 1999a,b).

Precyzję w warunkach odtwarzalności można ocenić poprzez obliczenie współczynnika zmienności V dla całego badanego zbioru danych ($N = 66$), przyjmując że warunki podlegające zmianom to:

- metodyka pomiarów (metoda AAS — wyniki oznaczeń w podzbiorach I–III; metoda ICP-AES — wyniki w podzbiorach IV–V);
- metodyka opróbowania (sprzęt wielokrotnego użytku Eijkelkamp — podzbiory I–II; sprzęt jednorazowego użytku firmy Millipore — podzbiory III–V);
- czas wykonywania pomiarów (lata: 1993, 1996, 1998/1999, 1999/2000).

Precyzja w warunkach odtwarzalności (zmiana metodyki oznaczeń, zmiany operatorów pobierających próbki, różny czas wykonywania badań) jest bardzo niska (tab. 1.8), o czym świadczy współczynnik zmienności obliczony dla całego zbioru danych $V = 322.22\%$.

O precyzji oznaczeń w warunkach odtwarzalności można też wnioskować na podstawie wartości współczynnika zmienności V obliczonego dla dwóch grup wyników pomiarów wydzielonych ze względu na różnice w metodyce oznaczeń ($N_1 = 29$ — wyniki oznaczeń w podzbiorach I–III, metoda oznaczania cynku: AAS; $N_2 = 37$ — wyniki w podzbiorach IV–V, metoda oznaczania cynku: ICP AES). Czynniki podlegające zmianie w ramach każdej z grup to operatorzy pobierający próbki, metodyka opróbowania i czas wykonywania oznaczeń.

Precyzja w warunkach odtwarzalności w obu grupach wydzielonych na podstawie różnic w metodyce badań (AAS, ICP-AES) jest bardzo niska ($V = 217.44\%$, $V = 173.09\%$). Lepsze wyniki uzyskano jednak dla oznaczeń cynku metodą ICP-AES o niższej granicy oznaczalności DL, niż metodą AAS (tab. 1.9).

Tabela 1.8. Ocena precyzji oznaczeń cynku [mg/dm^3] w próbkach wody jurajskiej ze Zdroju Królewskiego w Krakowie, w warunkach odtwarzalności

Statystyka opisowa	Pełny zbiór danych
Liczba pomiarów N	66
Rozstęp R	0.490
Minimum $x_{\min}$	0.010
Maksimum $x_{\max}$	0.500
Wartość średnia $\bar{x}$	0.0198
Odchylenie standardowe s	0.0638
Wariancja s^2	0.00007
Współczynnik zmienności $V = s/\bar{x}$	3.2222

Tabela 1.9. Ocena precyzji oznaczeń cynku [mg/dm^3] w próbkach wody jurajskiej ze Zdroju Królewskiego w Krakowie, w warunkach odtwarzalności. Metoda AAS: I — Tarnów 1993, II — Katowice 1993, III — Tarnów 1996; metoda ICP-AES: IV — ZHiOW AGH 1998/1999, V — ZHiOW AGH 1999/2000

Statystyka opisowa	Podzbiory: I+II+III		Podzbiory: IV+V	
Liczba pomiarów N	29	28	37	35
Granica oznaczalności DL	0.01	0.01	0.005	0.005
Rozstęp R	0.490	0.135	0.133	0.045
Minimum $x_{\min}$	0.010	0.005	0.005	0.005
Maksimum $x_{\max}$	0.500	0.145	0.138	0.050
Wartość średnia $\bar{x}$	0.0422	0.0258	0.0142	0.0093
Odchylenie standardowe s	0.0917	0.0261	0.0246	0.0098
Wariancja s^2	0.00841	0.00068	0.00061	0.00009
Współczynnik zmienności $V = s/\bar{x}$	2.1744	1.0099	1.7309	1.0518

Po wyłączeniu z analizy obserwacji leżących ponad górną granicą kontrolną (próbki nr 9, 31 i 47) precyzja oznaczeń ulega znacznej poprawie (tab. 1.9), niemniej jednak współczynnik zmienności nadal ma bardzo wysoką wartość ($V = 100.99\%$, $V = 105.18\%$).

Ocena precyzji oznaczeń cynku w próbkach wody jurajskiej ze Zdroju Królewskiego, w warunkach powtarzalności

W celu oszacowania precyzji oznaczeń w warunkach powtarzalności, dla każdego wydzielonego podzbioru danych I–V wyznaczono podstawowe statystyki opisowe (tab. 1.10). W każdym z podzbiorów była zastosowana ta sama metodyka pobierania i analizy próbek, proces opróbowania prowadzony był przez jednego operatora.

Wyniki uzyskane w laboratorium WIOŚ w Tarnowie w 1993 roku (podzbiór I) cechują się niską precyzją, bowiem współczynnik $V = 190.86\%$. Również duży współczynnik zmienności cechuje wyniki uzyskane w laboratorium ZHiOW AGH w latach 1999/2000 (podzbiór V, $V = 160.20\%$).

Po wyłączeniu z analizy wyników dla próbek nr 9, 30–34 i 45–50 (tab. 1.11) precyzja ulega poprawie. Fakt, że w laboratorium OBiKŚ w Katowicach (podzbiór II) analizowano tylko trzy próbki uniemożliwia wiarygodną ocenę precyzji.

Najlepsza precyzja oznaczeń cechuje wyniki uzyskane w 1996 roku, w laboratorium WIOŚ Tarnów (podzbiór III, $V = 20.06\%$). Dla badań prowadzonych w latach 1998–2000 w laboratorium ZHiOW AGH (podzbiory IV i V) — współczynnik zmienności $V \approx 50\%$. W obu przypadkach opróbowanie prowadzono z wykorzystaniem sprzętu jednorazowego

użytku Millipore. Dla badań przeprowadzonych w roku 1993, przy zastosowaniu sprzętu wielokrotnego użytku, precyzja oznaczeń cynku zarówno w warunkach powtarzalności, jak i odtwarzalności jest niezadowalająca.

Tabela 1.10. Ocena precyzji oznaczeń cynku [mg/dm^3] w próbkach wody jurajskiej ze Zdroju Królewskiego w Krakowie, w warunkach powtarzalności. Metoda AAS: I — Tarnów 1993, II — Katowice 1993, III — Tarnów 1996; metoda ICP-AES: IV — ZHiOW AGH 1998/1999, V — ZHiOW AGH 1999/2000

Statystyka opisowa	I	II	III	IV	V
Liczba pomiarów N	17	3	9	15	22
Granica oznaczalności DL	0.01	0.01	0.01	0.005	0.005
Rozstęp R	0.490	0.003	0.007	0.011	0.133
Minimum $x_{\min}$	0.010	0.023	0.010	0.005	0.005
Maksimum $x_{\max}$	0.500	0.026	0.017	0.016	0.138
Wartość średnia $\bar{x}$	0.0614	0.0243	0.0118	0.0067	0.0194
Odchylenie standardowe s	0.1172	0.0015	0.0028	0.0033	0.0310
Wariancja s^2	0.013740	0.000002	0.000008	0.000011	0.000962
Współczynnik zmienności $V = s/\bar{x}$	1.9086	0.0627	0.2430	0.4899	1.6020

Tabela 1.11. Ocena precyzji oznaczeń cynku [mg/dm^3] w próbkach wody jurajskiej ze Zdroju Królewskiego w Krakowie, w warunkach powtarzalności. Dane po wyłączeniu z analizy próbek: 9, 29, 30, 31, 33, 47 i 50. Metoda AAS: I — Tarnów 1993, II — Katowice 1993, III — Tarnów 1996; metoda ICP-AES: IV — ZHiOW AGH 1998/1999, V — ZHiOW AGH 1999/2000

Statystyka opisowa	I	II	III	IV	V
Liczba pomiarów N	16	3	8	15	17
Granica oznaczalności DL	0.01	0.01	0.01	0.005	0.005
Rozstęp R	0.135	0.003	0.006	0.011	0.014
Minimum $x_{\min}$	0.010	0.023	0.010	0.005	0.005
Maksimum $x_{\max}$	0.145	0.026	0.016	0.016	0.019
Wartość średnia $\bar{x}$	0.0340	0.0243	0.0111	0.0067	0.0066
Odchylenie standardowe s	0.0321	0.0153	0.0022	0.0003	0.0037
Wariancja s^2	0.001029	0.000002	0.000005	0.000011	0.000014
Współczynnik zmienności $V = s/\bar{x}$	0.9436	0.0628	0.2006	0.4899	0.5602

Precyzja oznaczeń cynku w warunkach powtarzalności (liczona dla każdego laboratorium z osobna) jest co najmniej 2–3-krotnie wyższa ($V \approx 20$–50%) niż precyzja wyznaczana w warunkach odtwarzalności ($V \approx 100\%$), co jest zgodne z danymi literaturowymi (Huber, 1997). Po wyłączeniu z obliczeń wyników obarczonych błędami grubymi, precyzja oznaczeń ulega znacznej poprawie. Obliczenia przeprowadzone dla próbek analizowanych w WIOŚ Tarnów w 1993 roku w warunkach powtarzalności (tab. 1.9) wskazują, że wyłączenie z obliczeń tylko jednej próbki obarczonej błędem grubym powoduje zmianę V z 190.86 do 94.36%.

W literaturze (Huber, 1997) zaleca się, by dla badań wykonywanych jednego dnia, badań kilkudniowych czy badań międzylaboratoryjnych, wartość współczynnika zmienności V nie przekraczała 20%. W przypadku badań prowadzonych w roku 1996 i w latach 1998/1999 i 1999/2000 (III–V) proces opróbowania wody jurajskiej prowadzony był przez jednego operatora, z wykorzystaniem sprzętu do filtracji i opróbowania jednorazowego użytku Millipore, co pozwoliło na redukcję błędów związanych z tym etapem badań. Współczynnik zmienności osiąga tutaj mniejszą wartość $V \approx 20$–50% w porównaniu do opróbowania prowadzonego z wykorzystaniem sprzętu wielokrotnego użytku Eijkelkamp (podzbiory I–II), gdzie $V \approx 190\%$.

Dane literaturowe (Thompson, Howarth, 1976) wskazują także, że precyzja oznaczeń danego składnika maleje, gdy mierzone stężenia są bliskie granicy oznaczalności DL. Dopuszczalny poziom precyzji ($V \leqslant 20\%$) uzyskuje się zazwyczaj, gdy mierzone stężenia są co najmniej o jeden rząd wielkości wyższe od granicy oznaczalności DL dla zastosowanej metody

pomiarowej. Zastosowanie zatem w badaniach metody pomiarowej o granicy oznaczalności o jeden rząd wielkości niższej (metoda ICP-AES zamiast AAS) wpływa istotnie na zwiększenie precyzji uzyskiwanych oznaczeń. Jeśli jednak mierzone stężenia są zbliżone do granicy oznaczalności DL, wówczas precyzja wyników może być niezadowalająca zarówno w warunkach powtarzalności, jak i odtwarzalności.

Z danych literaturowych (Nielsen, 1991) wynika również, że ok. 90% błędów w monitoringu wód podziemnych jest generowane na etapie procesu opróbowania. Analiza przeprowadzona dla cynku potwierdziła znaczący udział tego procesu w generowaniu błędów grubych, wpływających bezpośrednio na obniżenie precyzji wyników oznaczeń tego składnika.

Do niekontrolowanego wzrostu stężeń składnika w analizowanych próbkach (błędy przypadkowe, grube) może też prowadzić niewłaściwie prowadzona filtracja próbek wody. W efekcie przyczynia się to do zwiększenia rozrzutu wyników, a zatem obniżenia precyzji.

Aby uzyskiwać wiarygodne wyniki oznaczeń wskaźników jakości wód podziemnych należy tak planować badania monitoringowe, by próbki wody pobierane były przy zastosowaniu **sprzętu jednokrotnego użytku**, filtrowane *on-line* w terenie, a następnie analizowane **w jednym laboratorium**, **w warunkach powtarzalności pomiarów**, przy zastosowaniu **metody pomiarowej o odpowiednio niskiej granicy oznaczalności** DL w stosunku do spodziewanych stężeń w wodzie podziemnej.

1.2. Analiza rozkładu wskaźników fizyko-chemicznych wód podziemnych dorzecza górnej Wisły

Wyniki badań prowadzonych w ramach pierwszej serii opróbowania, w sieci regionalnego monitoringu jakości wód podziemnych RMWP dorzecza górnej Wisły (Witczak et al., 1994) pozwalają na uzyskanie obrazu stanu jakości wód podziemnych w zlewni górnej Wisły.

W serii tej opróbowaniem objęto 167 punktów RMWP (sieć składa się ze 172 punktów ale ze względu na niezakończony proces adaptacji punkty: 11012, 21024, 21047, 21052, 21060 nie zostały opróbowane), w próbkach wody oznaczano maksymalnie 55 cech i wskaźników (55 zmiennych). Ponieważ nie we wszystkich próbkach oznaczano wszystkie deklarowane wskaźniki, zmienia się liczba danych (obserwacji) w poszczególnych zmiennych — braki danych.

Opis statystyczny analizowanej bazy danych (składającej się z 55 zmiennych — wskaźników fizyko-chemicznych wód i 167 obserwacji — opróbowanych punktów RMWP) uzyskano za pomocą programu SPSS PL for Windows v. 10.0. Wykorzystano do tego celu procedurę eksploracji zbioru danych:

menu: **Analiza ▶ Opis statystyczny ▶ Eksploracja**

oraz kart kontrolnych:

menu: **Wykresy ▶ Karty kontrolne**.

Wyniki oznaczeń wskaźników chemicznych poniżej granicy oznaczalności zostały przyjęte do obliczeń jako $< \mathrm{DL} = \mathrm{DL}$ (Helsel, Hirsch, 1992). W przypadku gdy wyniki $< \mathrm{DL}$ stanowią ponad 20% obserwacji danej zmiennej, uzyskany w wyniku analizy statystycznej rozkład tej zmiennej jest zniekształcony przez te obserwacje (Górniak, Wachnicki, 2000).

Przeprowadzono analizę rozkładu wskaźników fizyko-chemicznych wód, zidentyfikowano wartości oznaczone na wykresach typu „skrzynka z wąsami" jako ekstremalne (Luszniewicz, Słaby, 1998). Obserwacje te pokrywały się z obserwacjami ekstremalnymi z wykresów typu „łodyga i liście". Na tej podstawie dokonano podziału badanego zbioru analiz na podzbiory, charakteryzujące subpopulacje: anomalną (obserwacje ekstremalne) i typową — obserwacje typowe, pozostałe w zbiorze po wyłączeniu z analizy obserwacji ekstremalnych (Macioszczyk, 1990; Siwek, 1999).

Następnie ponownie wykonano analizę rozkładu badanych zmiennych dla subpopulacji typowej, i w niektórych przypadkach — gdy liczba próbek wyłączonych z analizy była

większa od siedmiu, analizę opisową subpopulacji anomalnej, wykorzystując do tego celu procedury eksploracji i częstości w programie SPSS PL for Windows.

W dodatkach C, D i E znajdują się przykłady raportów z analizy rozkładu pełnego zbioru danych, podzbioru wartości typowych i podzbioru wartości anomalnych dla cynku. Pełny zestaw raportów z analiz przeprowadzonych dla wszystkich badanych wskaźników jakości wód znajduje się na płycie CD-ROM dołączonej do niniejszej pracy.

W ramach procedury eksploracji analizowanych populacji i subpopulacji uzyskano:

- informacje ogólne o analizowanych zmiennych (rys. 1.17) obejmujące ogółem liczbę obserwacji w zbiorze (*Ogółem*), liczbę obserwacji analizowanych (*Uwzględnione*) i liczbę wyłączonych z analizy braków danych (*Wykluczone*);

Informacja o analizowanych danych

	Obserwacje					
	Uwzględnione		Wykluczone		Ogółem	
	N	Procent	N	Procent	N	Procent
Cynk [mg/dm3]	166	99.4%	1	.6%	167	100.0%

Rysunek 1.17. Procedura eksploracji zbioru danych — informacje ogólne o zmiennej *cynk* [mg/dm^3]

- zestawienie statystyk opisowych (rys. 1.18);

Statystyki opisowe

		Statystyka	Błąd standardowy
Średnia		.23149	9.9891E-02
95% przedział ufności dla średniej	Dolna granica	3.4264E-02	
	Górna granica	.42872	
5% średnia obcięta		7.0948E-02	
Mediana		4.0000E-02	
Wariancja		1.656	
Odchylenie standardowe		1.28701	
Minimum		.003	
Maksimum		15.000	
Rozstęp		14.997	
Rozstęp ćwiartkowy		7.5500E-02	
Skośność		10.174	.188
Kurtoza		110.748	.375

Rysunek 1.18. Procedura eksploracji zbioru danych — zestawienie statystyk opisowych dla zmiennej *cynk* [mg/dm^3]

Zestawienie to obejmuje:

– miary położenia: wartość średnia z błędem standardowym i 95% przedziałem ufności, 5%-średnia obcięta (wartość średnia obliczona po obcięciu 5% obserwacji największych i najmniejszych w zbiorze, pozwala ocenić jaki wpływ na wartość średnią mają obserwacje odstające, jeśli wpływ ten jest znikomy, 5%-średnia obcięta jest bliska wartości średniej obliczonej dla całego zbioru danych), mediana;
– miary rozproszenia: wariancja, odchylenie standardowe, miminum, maksimum, rozstęp (różnica między największą i najmniejszą wartością w zbiorze), rozstęp ćwiartkowy (różnica pomiędzy wartościami pierwszego i trzeciego kwartyla);

– miary asymetrii: skośność, kurtoza, błąd standardowy skośności i kurtozy);[2]

- percentyle: 5, 10, 25 (pierwszy kwartyl), 50 (mediana), 75 (trzeci kwartyl), 90 i 95 (rys. 1.19); zawiasy Tukey'a wyznaczają kwartyle rozkładu;

Percentyle

Percentyle	Przeciętne ważone (Definicja 1)	Zawiasy Tukey'a
5	9.3500E-03	
10	1.0000E-02	
25	2.0000E-02	2.0000E-02
50	4.0000E-02	4.0000E-02
75	9.5500E-02	9.5000E-02
90	.25470	
95	.52685	

Rysunek 1.19. Procedura eksploracji zbioru danych — zestawienie percentyli dla zmiennej *cynk* [mg/dm^3]

- wartości skrajne — pięć najwyższych i pięć najniższych wartości w analizowanym zbiorze danych (rys. 1.20);

Wartości skrajne

		Numer obserwacji	Numer id. punktu w bazie MONBADA	Wartość
Najwyższe	1	13	11014	15.000
	2	12	11013	7.000
	3	114	21067	1.900
	4	60	21006	.824
	5	1	11001	.800
Najniższe	1	152	21105	.003
	2	151	21104	.004
	3	72	21020	.004
	4	108	21061	.005
	5	96	21046	.006

Rysunek 1.20. Procedura eksploracji zbioru danych — zestawienie wartości skrajnych dla zmiennej *cynk* [mg/dm^3]

- wykres typu łodyga i liście — charakteryzujący rozkład empiryczny badanego wskaźnika, pozwala wyznaczyć wartości ekstremalne; dzieli badany zbiór na klasy, „łodygę" tworzą początkowe cyfry wartości zmiennej uporządkowane rosnąco i oddzielone od „liści" znakiem kropki, „liście" tworzą kolejne cyfry zmiennej, przy czym każdy „liść" tworzą dwie pary cyfr, rozpoczynając od pary 0 i 1, a kończąc na parze 8 i 9; jeżeli rozstęp między wartością minimalną i maksymalną jest duży, procedura automatycznie zwiększa zakres „liścia" do pięciu kolejnych cyfr danego rzędu wielkości: od 0 do 4 i od 5 do 9; długość „liścia" zależy od liczby wartości zmiennej, w których cyfry są identyczne, np. zapis **8 . 116** oznacza, że wśród wartości zmiennej występują dwie wartości rozpoczynające się od cyfr **8, 1** i jedna wartość rozpoczynająca się od cyfr **8, 6** (rys. 1.21);
- wykres skrzynkowy — tworzony na podstawie mediany, kwartyli i wartości skrajnych (rys. 1.22);

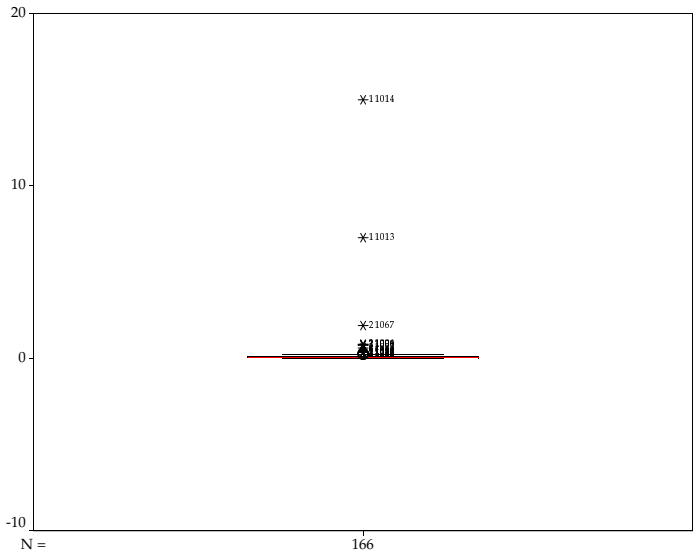
(2) Odpowiednie definicje i wzory dotyczące omawianych statystyk można znaleźć w dodatku B oraz w literaturze, np. Górniak, Wachnicki, 2000.

```
Cynk [mg/dm3] Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     8.00        0 .  34456899
    23.00        1 .  00000000000000012566699
    27.00        2 .  000000000000000111233456788
    23.00        3 .  00000000000111234667788
    12.00        4 .  000111233348
    10.00        5 .  0000223588
     9.00        6 .  000003688
     8.00        7 .  00233569
     3.00        8 .  116
     3.00        9 .  057
     4.00       10 .  5667
     1.00       11 .  6
     4.00       12 .  0025
     1.00       13 .  5
     2.00       14 .  39
     1.00       15 .  8
     3.00       16 .  006
     2.00       17 .  07
      .00       18 .
     2.00       19 .  15
    20.00 Extremes    (>=.220)

Stem width:      .010
Each leaf:       1 case(s)
```

Rysunek 1.21. Procedura eksploracji zbioru danych — wykres typu „łodyga i liście" dla zmiennej *cynk* [mg/dm^3]

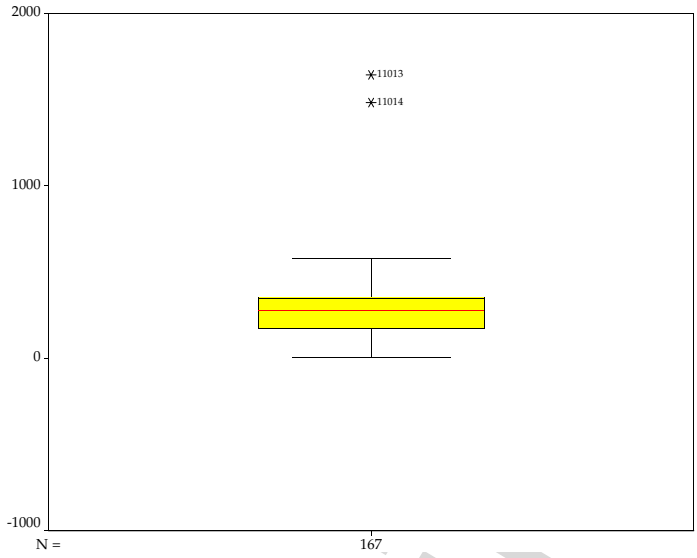
W przypadku cynku wykres ten jest mało czytelny; bardziej czytelną postać wykresu uzyskano np. dla twardości ogólnej (rys. 1.23).

Linię w środku skrzynki wyznacza mediana, góra i dół skrzynki (zawiasy, *hinges*) są wyznaczone przez wartości pierwszego i trzeciego kwartyla (długość skrzynki reprezentuje rozrzut środkowych 50% obserwacji, tzw. rozstęp ćwiartkowy). Wąsy (*whiskers*) — linie przyczepione do zawiasów reprezentują rozstęp wartości, które mieszczą się w zakresie 1.5 rozstępu ćwiartkowego od zawiasów. Obserwacje, które mają wartości leżące poza zasięgiem wąsów traktowane są jako obserwacje odstające, wpływające zbyt silnie na oszacowanie średniej i inne statystyki. Program wyświetla symbole graficzne oznaczające obserwacje odstające — kółka dla obserwacji odstających leżących w zakresie 1.5 do 3



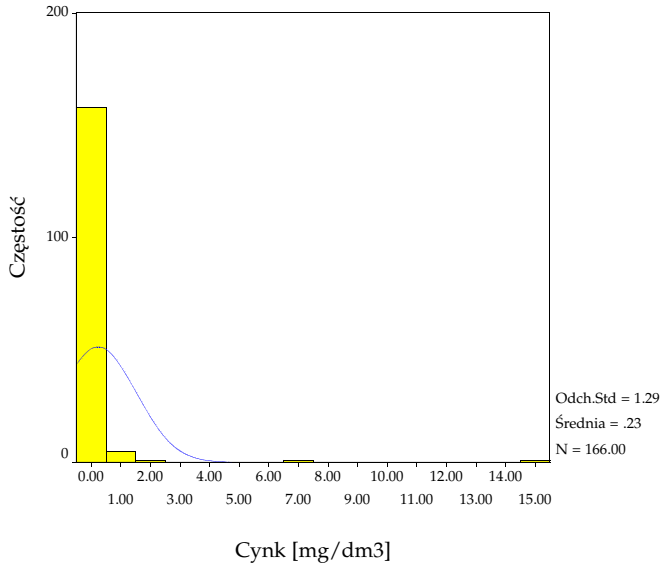
Rysunek 1.22. Procedura eksploracji zbioru danych — wykres skrzynkowy dla zmiennej *cynk* [mg/dm^3]

razy rozstęp ćwiartkowy od zawiasów, i gwiazdki dla obserwacji ekstremalnych, spoza granicy 3 rozstępów ćwiartkowych od zawiasów.



Rysunek 1.23. Procedura eksploracji zbioru danych — wykres skrzynkowy dla zmiennej *twardość ogólna* [mg/dm^3]; widoczne są dwie obserwacje ekstremalne

- histogram rozkładu pojedynczych pomiarów z krzywą gęstości rozkładu (rys. 1.24);



Rysunek 1.24. Procedura eksploracji zbioru danych — histogram rozkładu pojedynczych pomiarów z krzywą gęstości rozkładu dla zmiennej *cynk* [mg/dm^3]

- wynik testu normalności rozkładu (rys. 1.25);
  W kolumnie *Statystyka* znajduje się wartość statystyki Kołmogorowa-Smirnowa, w kolumnie *df* — liczba stopni swobody (*degrees of freedom*). Kolumna *Istotność* z poprawką Lillieforsa zawiera wielkość prawdopodobieństwa popełnienia błędu I rodzaju (błąd pierwszego rodzaju polega na odrzucenia hipotezy zerowej testu, gdy jest ona prawdziwa), w tym przypadku hipoteza brzmi: „nie ma podstaw do odrzucenia hipotezy, że analizowana zmienna ma rozkład normalny".

Testy normalności rozkładu

	Kołmogorow-Smirnow[a]		
	Statystyka	df	Istotność
Cynk [mg/dm3]	.430	166	.000

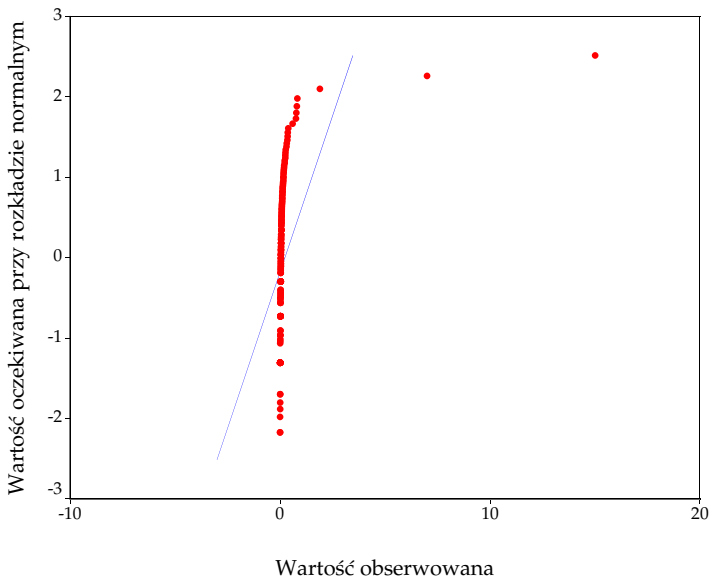
a. Z poprawką istotności Lillieforsa

Rysunek 1.25. Procedura eksploracji zbioru danych — wynik testu normalności rozkładu dla zmiennej *cynk* [mg/dm^3]

Hipotezę zerową testu odrzucamy, gdy wartość prawdopodobieństwa testowego jest mniejsza od 0.05 — w przypadku analizowanej zmiennej (cynk) prawdopodobieństwo testowe ma wartość 0.000, zatem należy odrzucić hipotezę zerową o normalności rozkładu tych danych.

Graniczną wartość tego prawdopodobieństwa, poniżej której odrzucamy hipotezę zerową, nazywamy poziomem istotności.

- wykres normalności rozkładu (rys. 1.26);

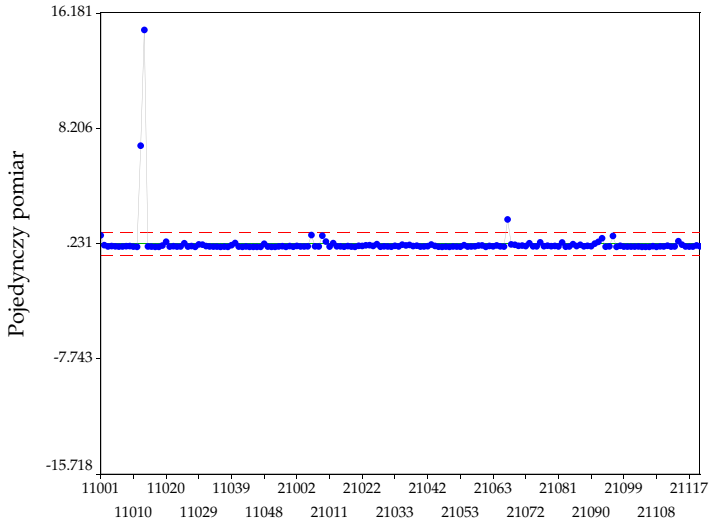


Rysunek 1.26. Procedura eksploracji zbioru danych — wykres normalności rozkładu dla zmiennej *cynk* [mg/dm^3]

Widoczna na wykresie prosta pozwala ocenić na ile rozkład wartości badanej zmiennej odbiega od rozkładu normalnego (w przypadku rozkładu normalnego punkty na wykresie układają się na tej prostej);

- kartę kontrolną pojedynczych pomiarów (rys. 1.27).

Karta kontrolna to wykres, na którym rejestruje się wartości kontrolowanego parametru (np. stężenie) w funkcji czasu pobierania próbki lub numeru próbki. Na karcie umieszczone są linie kontrolne: linia centralna i granice kontrolne. O jakości kontrolowanego parametru wnioskuje się na podstawie położenia na karcie punktów odpowiadających wartościom rejestrowanego parametru względem linii kontrolnych. W idealnym przypadku wszystkie punkty powinny leżeć wewnątrz przedziału ograniczonego granicami kontrolnymi i być równomiernie rozrzucone wokół linii centralnej. Szczegóły dotyczące rodzajów kart kontrolnych, ich konstrukcji i analizy można znaleźć w literaturze (Kmiecik, 1994; Szczepańska, Kmiecik, 1998).



Rysunek 1.27. Procedura eksploracji zbioru danych — karta kontrolna pojedynczych pomiarów dla zmiennej *cynk* [mg/dm^3]

W tabeli 1.12 zestawiono charakterystykę analizowanych zmiennych wraz ze współczynnikami zmienności obliczonymi na podstawie estymowanych parametrów rozkładu, w rozdziale 1.2 znajduje się krótki opis analizy rozkładu każdej ze zmiennych.

Opisowa analiza statystyczna badanych wskaźników fizyko-chemicznych

Wyniki badań fizyko-chemicznych wód podziemnych dorzecza górnej Wisły w obszarach RZGW Katowicz i RZGW Kraków stanowią realizacje zmiennych losowych, na które oddziałują przyczyny główne: wpływ ognisk zanieczyszczeń, procesy hydrogeochemiczne zachodzące w strefie aeracji i saturacji, trendy zmian jakości wód i przyczyny uboczne: technika poboru próbek, doświadczenie grupy terenowej, prawidłowość procedur, itp. (Bednarczyk, 1998; Siwek, 1999). Zmienne te poddane zostały analizie statystycznej, w celu określenia ich rozkładu i wykrycia obserwacji odstających (anomalnych).

Badania terenowe

Wskaźniki fizyko-chemiczne oznaczane w terenie, bezpośrednio przy poborze próbek są wskaźnikami nietrwałymi. Ich ocena jest w większości przypadków jakościowa (np. mętność, barwa czy zapach). Analizą statystyczną objęte zostały: temperatura, przewodność elektrolityczna właściwa, odczyn pH, potencjał utleniająco-redukcyjny Eh, zasadowość ogólna, zasadowość mineralna i kwasowość ogólna, natomiast w dalszej analizie uwzględnione zostaną tylko temperatura i pH.

Temperatura

Temperatura w badanym zbiorze 167 punktów RMWP zmieniała się w przedziale od 5.0 do 15.0°C. Wartość średnia wynosi $\bar{x} = 9.97$°C i jest bliska 5%-średniej obciętej, co oznacza, że na wartość średnią nie mają wpływu obserwacje odstające.

Tabela 1.12. Opis analizowanych wskaźników fizyko-chemicznych wód podziemnych w zlewni górnej Wisły. Objaśnienia: N — liczba obserwacji uwzględnionych, B — liczba braków danych (nie wszystkie wskaźniki były oznaczane w każdej z próbek), V — współczynnik zmienności

Lp.	Analizowana zmienna	Jednostka	N	B	V [%]
Oznaczenia terenowe					
1.	Temperatura	°C	167	0	16.19
2.	Przewodność	μS/cm	163	4	76.75
3.	Odczyn pH		167	0	9.73
4.	Potencjał redox Eh	mV	167	0	166.02
5.	Zasadowość ogólna	mval/dm^3	155	12	51.11
6.	Kwasowość ogólna	mval/dm^3	164	3	109.45
Oznaczenia laboratoryjne					
1.	Suma substancji rozpuszczonych	mg/dm^3	166	1	80.65
2.	Zasadowość ogólna	mval/dm^3	167	0	47.65
3.	Twardość ogólna	mg $CaCO_3$/dm^3	167	0	65.92
4.	Potas	mg/dm^3	166	1	143.89
5.	Sód	mg/dm^3	166	1	202.17
6.	Magnez	mg/dm^3	167	0	121.50
7.	Wapń	mg/dm^3	167	0	59.35
8.	Azot amonowy	mg/dm^3	167	0	184.20
9.	Glin	mg/dm^3	166	1	89.71
10.	Żelazo ogólne	mg/dm^3	167	0	284.99
11.	Mangan	mg/dm^3	167	0	653.91
12.	Azot azotynowy	mg/dm^3	167	0	181.61
13.	Azot azotanowy	mg/dm^3	167	0	127.82
14.	Chlorki	mg/dm^3	167	0	147.56
15.	Siarczany	mg/dm^3	167	0	209.47
16.	Fosforany rozpuszczone	mg/dm^3	166	1	139.74
17.	Krzemionka zdysocjowana	mg/dm^3	166	1	65.97
18.	Fluorki	mg/dm^3	162	5	80.75
19.	Bor	mg/dm^3	41	126	201.46
20.	Chrom ogólny	mg/dm^3	166	1	102.92
21.	Cynk	mg/dm^3	166	1	555.97
22.	Kadm	mg/dm^3	164	3	248.72
23.	Miedź	mg/dm^3	165	2	74.36
24.	Nikiel	mg/dm^3	166	1	133.76
25.	Ołów	mg/dm^3	164	3	89.61
26.	Rtęć	mg/dm^3	166	1	106.94
27.	Współczynnik absorpcji UV (A254)		167	0	191.17
28.	Rozpuszczony węgiel organiczny	mg/dm^3	164	3	103.46
29.	Utlenialność ChZT-Mn	mg/dm^3	167	0	97.08
30.	Azot organiczny Kjeldahla	mg/dm^3	166	1	134.36
31.	Fenole lotne	mg/dm^3	166	1	228.39
32.	Substancje ropopochodne	mg/dm^3	166	1	138.39
33.	Chloroform	mg/dm^3	166	1	236.40
34.	Subst. pow.-czynne anionowe	mg/dm^3	166	1	43.39
35.	Czterochloroetylen	mg/dm^3	166	1	223.81
36.	Trójchloroetylen	mg/dm^3	166	1	589.03
37.	DDT	mg/dm^3	166	1	56.29
38.	DDE	mg/dm^3	166	1	24.72
39.	DDD	mg/dm^3	166	1	116.31
40.	Gamma-HCH	mg/dm^3	166	1	36.47
41.	Metoksychlor	mg/dm^3	166	1	123.70
42.	Benzo-a-piren	mg/dm^3	167	0	—
43.	Suma 6WWA	mg/dm^3	53	114	62.21

— nie obliczano współczynnika zmienności, gdyż wszystkie obserwacje były $<$ DL

Rozkład charakteryzuje się lekką asymetrią prawostronną (współczynnik skośności wynosi 0.155) i jest spiczasty (kurtoza: 1.090). Współczynnik zmienności dla badanego zbioru wynosi $V = 16.19\%$.

Wykres typu „łodyga i liście" wykazuje istnienie czterech obserwacji anomalnie niskich $\leqslant 6.0^{\circ}$C (punkty: 11052, 11048, 11053, 21066)[3] oraz trzech obserwacji anomalnie wysokich $\geqslant 14.0^{\circ}$C (punkty: 21008, 11033, 21073).

Po wyłączeniu z analizy obserwacji odstających wartość średnia wzrosła do 9.99°C, przy rozrzucie wartości od 6.5 do 13.4°C. Rozkład jest w przybliżeniu rozkładem normalnym, współczynnik zmienności $V = 13.6\%$.

Przewodność elektrolityczna właściwa

Wartość tego parametru zmieniała się w zakresie 63–3780 μS/cm, przy wartości średniej $\bar{x} = 519.42$ μS/cm. 5%-średnia obcięta wynosi 479.45 μS/cm, co oznacza, że na wartość średnią wpływają wartości odstające. Rozkład charakteryzuje się asymetrią prawostronną (skośność: 4.920) i spiczastością (kurtoza: 35.323). Współczynnik zmienności zbioru $V = 76.75\%$.

Na wykresie typu „łodyga i liście" widoczne jest sześć obserwacji ekstremalnie wysokich (punkty: 11029, 21096, 11019, 21053, 11014, 11013).

Odczyn pH

Analizowane wody podziemne charakteryzują się odczynem od kwaśnego po słabo zasadowy (pH = 4.6 – 8.7). Najbardziej typowe wartości (95% przedział ufności dla średniej) mieściły się w przedziale od 6.94 do 7.15 (wartość średnia: $\bar{x} = 7.04$, 5%-średnia obcięta: 7.08). Rozkład charakteryzuje się asymetrią lewostronną (skośność: -1.070) i spiczastością (kurtoza: 1.716). Współczynnik zmienności ma niską wartość $V = 9.73\%$.

W badanym zbiorze występuje dziesięć wartości ekstremalnie niskich $x \leqslant 5.6$ i jedna wartość ekstremalnie wysoka $x \geqslant 8.7$.

Po wyłączeniu z analizy obserwacji odstających rozkład pH jest w przybliżeniu rozkładem normalnym. Parametr ten zmienia się w zakresie 5.8–8.4, przy wartości średniej $\bar{x} = 7.15$, ze współczynnikiem zmienności $V = 7.04\%$.

Potencjał utleniająco-redukcyjny Eh

W badanym zbiorze potencjał utleniająco-redukcyjny Eh zmienia się w zakresie od -241 mV do 1000 mV. Wartość średnia wynosi $\bar{x} = 80.39$ mV, a 5%-średnia obcięta 80.35 mV, co oznacza brak wpływu na średnią obserwacji odstających. Rozkład charakteryzuje się asymetrią prawostronną (współczynnik skośności: 1.492) i znaczną spiczastością (kurtoza: 12.589). Współczynnik zmienności ma dużą wartość $V = 166.02\%$ (w przypadku tej zmiennej istniały trudności z precyzyjnym pomiarem, jest to jedynie ocena przybliżona — Witczak et al., 1994).

W zbiorze tym są dwie obserwacje ekstremalnie niskie $x \leqslant -212$ mV (punkty: 21085, 11037) i jedna ekstremalnie wysoka — punkt 21021 ($x \geqslant 1000$ mV).

Mętność

Ze względu na różnice w metodyce oznaczania tego parametru pomiędzy RZGW Katowice (pomiar w [mg SiO_2/dm^3]) a RZGW Kraków (ocena jakościowa, opisowa) nie będzie prowadzona analiza statystyczna zbioru danych, a zmienna zostanie wyłączona z dalszej analizy.

(3) Numeracja punktów z bazy MONBADA.

Osad w leju Imhoffa (zawiesiny)

W próbkach z obszaru RZGW Katowice zmienna ta nie była oznaczana, w punktach z obszaru RZGW Kraków we wszystkich przypadkach ($N = 112$, jeden brak danych) przyjmuje stałe wartości 0 mg/dm^3. Nie będzie zatem prowadzona w tym przypadku analiza statystyczna zbioru danych i zmienna zostanie wyłączona z dalszej analizy.

Barwa

W obszarze RZGW Kraków barwa oceniana była jakościowo, w 93 przypadkach na 113 (jeden brak danych) roztwór był bezbarwny, w 18 przypadkach zaobserwowano delikatne żółte zabarwienie, a w 1 przypadku — białe (może to być spowodowane ługowaniem substancji organicznych, głównie związków humusowych; Witczak, Adamczyk, 1995).

W obszarze RZGW Katowice barwę oznaczano kolorymetrycznie — zmienna ta w 43 przypadkach przyjmuje wartość 0 mg Pt/dm^3, w pozostałych przypadkach osiąga wartości z zakresu 5–40 mg Pt/dm^3. Nie będzie zatem prowadzona dalsza analiza tej zmiennej.

Zapach

Zapach należy do cech wody oznaczanych organoleptycznie. W badanym zbiorze 167 punktów RMWP wystąpiły 73 braki danych, w 41 przypadkach nie odnotowano zapachu, w 46 przypadkach ujawnił się bardzo słaby zapach, najczęściej pochodzenia roślinnego (w 7 przypadkach był to zapach pochodzenia gnilnego, a w 9 przypadkach zapach specyficzny). W dwóch przypadkach odnotowano słaby zapach pochodzenia roślinnego, w 1 przypadku wyraźny zapach, również pochodzenia roślinnego, a w 4 przypadkach bardzo silny zapach specyficzny (punkty: 11009, 110021, 11034, 11035).

Zapachy z grupy zapachów roślinnych mogą być związane ze starzeniem się przewodów tłocznych lub obecnością niewielkiej ilości zawiesiny wodorotlenków żelaza w wodzie (Siwek, 1999).

Ze względu na charakter jakościowy zmienna ta będzie wyłączona z dalszej analizy.

Zasadowość ogólna

Zasadowość jest określana jako zdolność wody do zobojętniania silnego kwasu. Jest to cecha odwrotna do kwasowości. Przy wodach o pH < 8.3 (jak w analizowanym przypadku) zasadowość wynika z obecności wodorowęglanów HCO_3^- i jest podstawą do określania zawartości tego jonu.

W badanym zbiorze zasadowość ogólna zmienia się w zakresie 0.10–9.90 mval/dm^3. Wartość średnia wynosi $\bar{x} = 3.985$ mval/dm^3 a 5%-średnia obcięta: 3.952 mval/dm^3. Rozkład jest prawie symetryczny (skośność: 0.110, kurtoza: -0.467). Współczynnik zmienności wynosi $V = 51.11\%$.

Z wykresu typu „łodyga i liście" wynika, że w zbiorze jest jedna obserwacja odstająca (punkt 21011).

Zasadowość mineralna

Ponieważ pH w badanym zbiorze punktów mieści się w zakresie 4.6–8.7 zasadowość mineralna we wszystkich przypadkach ma wartość 0 mval/dm^3. Nie będzie zatem prowadzona analiza statystyczna tego zbioru danych.

Kwasowość ogólna

Ponieważ pH w analizowanym zbiorze danych jest większe od 4.5, więc kwasowość jest tu wywołana głównie przez dwutlenek węgla rozpuszczony w wodzie. Kwasowość ogólna mieści się w zakresie 0.01–5.85 mval/dm^3, przy wartości średniej 0.649 mval/dm^3.

Rozkład tej zmiennej charakteryzuje się asymetrią prawostronną (skośność: 3.293) i znaczną spiczastością (kurtoza: 17.849). Współczynnik zmienności ma wartość $V = 109.45\%$.

Na wykresie typu „łodyga i liście" widoczne jest dziesięć obserwacji odstających (punkty, w których kwasowość ogólna $\geqslant 1.89$ mval/dm^3).

Badania laboratoryjne — wskaźniki podstawowe i dodatkowe

Suma substancji rozpuszczonych

Parametr ten zmienia się w zakresie 46–2953 mg/dm^3, wartość średnia $\bar{x} = 378.15$ mg/dm^3 a 5%-średnia obcięta 347.71 mg/dm^3. Rozkład wykazuje asymetrię prawostronną (skośność: 5.513) i znaczną spiczastość (kurtoza: 41.580). Współczynnik zmienności wynosi $V = 80.65\%$.

W zbiorze są trzy obserwacje odstające (punkty: 11013, 11014, 11019). Po ich wyłączeniu z analizy zmienna ta ma rozkład w przybliżeniu normalny, z wartością średnią $\bar{x} = 346.82$ mg/dm^3 i współczynnikiem zmienności $V = 46.26\%$.

Wartości oznaczeń sumy substancji rozpuszczonych są dwa rzędy wielkości wyższe od granic oznaczalności deklarowanych przez oba laboratoria (tab. 1.3). Z danych literaturowych (Thompson, Howarth, 1976; Szczepańska, Kmiecik, 1998) wynika, że przy stężeniach wskaźników w wodach podziemnych o 1–2 rzędy wielkości wyższych od DL uzyskuje się odpowiedni poziom precyzji.

Praktyczna granica oznaczalności PDL jest czterokrotnie większa od laboratoryjnej granicy oznaczalności DL. Wariancja techniczna σ^2_{tech} (tab. 1.5) stanowi 5% zmienności całkowitej, co potwierdza zadowalającą precyzję oznaczeń tego wskaźnika.

Zasadowość ogólna

Zasadowość ogólna zmienia się w zakresie od 0.20 do 9.85 mval/dm^3, przy wartości średniej $\bar{x} = 4.063$ mval/dm^3 i 5%-średniej obciętej 4.037 mval/dm^3. Nie ma podstaw do odrzucenia hipotezy o normalności rozkładu wartości stężeń tego wskaźnika (skośność: 0.102, kurtoza: -0.317). W zbiorze jest jedna obserwacja odstająca (punkt 21096), współczynnik zmienności wynosi $V = 47.65\%$.

Po wyłączeniu z analizy obserwacji odstających wartość średnia nieznacznie zmalała $\bar{x} = 4.03$ mval/dm^3. Rozkład wartości jest normalny, z rozrzutem w granicach od 0.2 do 8.5 mval/dm^3 i współczynnikiem zmienności $V = 46.89\%$.

Wyniki oznaczeń są prawie dwa rzędy wielkości wyższe od deklarowanych przez laboratoria granic oznaczalności (tab. 1.3), zatem powinny cechować się zadowalającą precyzją.

Potwierdzeniem tego jest wariancja techniczna σ^2_{tech}, stanowiąca w tym przypadku zaledwie 4% zmienności całkowitej (tab. 1.5). Stosunek PDL/DL = 3.

Twardość ogólna

Parametr ten jest wyrazem obecności w wodzie metali ziem alkaicznych, przede wszystkim wapnia i magnezu oraz baru i strontu (Witczak, Adamczyk, 1995). Twardość ogólna badanych wód zmienia się od 3.60 mg $CaCO_3$/dm^3 do 1641.60 mg $CaCO_3$/dm^3 (odpowiada to wodom od bardzo miękkich po bardzo twarde), przy wartości średniej $\bar{x} = 283.80$ mg $CaCO_3$/dm^3 (wody średnio twarde) i 5%-średniej obciętej 269.44 mg $CaCO_3$/dm^3. Współczynnik zmienności wynosi $V = 65.92\%$.

Rozkład wartości tego parametru charakteryzuje się asymetrią prawostronną (skośność: 3.860) i spiczastością (kurtoza: 25.259). W zbiorze są dwie obserwacje odstające (punkty: 11013, 11014). Po ich wyłączeniu z analizy rozkład jest w przybliżeniu normalny, współczynnik zmienności ma wartość $V = 45.94\%$, a średnia $\bar{x} = 268.32$ mg $CaCO_3$/dm^3, przy zmienności badanego parametru w zakresie 3.60–575.51 mg $CaCO_3$/dm^3.

Wyniki oznaczeń twardości ogólnej cechują się zadowalającą precyzją, gdyż są dwa rzędy wielkości wyższe od granicy oznaczalności DL deklarowanej przez laboratoria wykonujące

analizy (tab. 1.3) a wariancja techniczna σ^2_{tech} stanowi zaledwie 3.61% zmienności całkowitej (tab. 1.5). Praktyczna granica oznaczalności PDL jest równa laboratoryjnej granicy oznaczalności DL.

Twardość węglanowa

Wartości twardości węglanowej uzyskane zostały na drodze obliczeniowej (pozostają w korelacji z wynikami oznaczeń twardości ogólnej), zatem wykonana zostanie analiza statystyczna tego zbioru, jednak zmienna ta nie będzie wykorzystana do prognozowania zmian jakości wód.

Wartość średnia twardości węglanowej wynosi $\bar{x} = 201.44$ mg $CaCO_3/dm^3$, przy rozrzucie wartości od 20.0 do 492.9 mg $CaCO_3/dm^3$. Nie ma podstaw do odrzucenia hipotezy o normalności rozkładu tej zmiennej (skośność: 0.115, kurtoza: -0.219). W badanym zbiorze jest jedna obserwacja odstająca (punkt 21096), współczynnik zmienności ma wartość $V = 45.65\%$.

Po wyłączeniu z analizy obserwacji odstających wartość średnia wynosi $\bar{x} = 199.98$ mg $CaCO_3/dm^3$, przy rozrzucie wartości od 20.00 do 425.37 mg $CaCO_3/dm^3$. Rozkład jest w przybliżeniu rozkładem normalnym, współczynnik zmienności zbioru $V = 44.76\%$.

Wyniki powinny cechować się zadowalającą precyzją, gdyż wariancja techniczna σ^2_{tech} stanowi 4% zmienności całkowitej.

Potas

Stężenia tego wskaźnika w analizowanych wodach podziemnych mieszczą się w zakresie od 0.1 mg/dm^3 do 21.1 mg/dm^3, przy wartości średniej $\bar{x} = 2.11$ mg/dm^3 i 5%-średniej obciętej 1.60 mg/dm^3. Rozkład cechuje się dużą asymetrią prawostronną (skośność: 3.664) i znaczną spiczastością (kurtoza: 15.792). Współczynnik zmienności ma wartość $V = 143.89\%$.

Na wykresie typu „łodyga i liście" widoczne jest szesnaście obserwacji odstających (stanowią one 9.64% obserwacji).

Obserwacje odstające zostały włączone z analizy — tworzą one podzbiór obserwacji anomalnych (Macioszczyk, 1990). Średnie stężenie potasu w tym podzbiorze $\bar{x} = 9.76$ mg/dm^3, przy rozrzucie wartości od 5.00 do 21.10 mg/dm^3.

W zbiorze wartości pozostałych (podzbiór wartości typowych) średnie stężenie potasu kształtuje się na poziomie $\bar{x} = 1.29$ mg/dm^3. Zmienna ta ma rozkład logarytmiczno-normalny, obserwacje mieszczą się w przedziale 0.1–4.4 mg/dm^3. Współczynnik zmienności zbioru $V = 72.76\%$.

Wyniki oznaczeń potasu są dwa rzędy wielkości wyższe od deklarowanych przez laboratoria granic oznaczalności, zatem powinny cechować się zadowalającą precyzją (tab. 1.3).

Praktyczna granica oznaczalności PDL jest jednak 150 razy wyższa od laboratoryjnej granicy oznaczalności, a wariancja techniczna wyznaczona metodą analizy wariancji stanowi aż 50% zmienności całkowitej (tab. 1.5). Zmienna ta będzie wyłączona z dalszej analizy.

Sód

W zbiorze wyników oznaczeń sodu w próbkach pobranych z punktów RMWP obszaru RZGW Kraków średnie stężenie kształtuje się na poziomie $\bar{x} = 13.08$ mg/dm^3, przy rozrzucie wartości od 0.5 do 244.8 mg/dm^3. Rozkład jest spiczasty (kurtoza: 42.141), z asymetrią prawostronną (skośność: 5.864). Współczynnik zmienności wynosi $V = 202.17\%$.

Wśród analizowanych stężeń sodu 6.63% obserwacji (jedenaście obserwacji) to obserwacje ekstremalne (wykres „łodyga i liście"). Są to wartości z przedziału 30.00–244.80 mg/dm^3, wartość średnia $\bar{x} = 88.78$ mg/dm^3.

Średnie stężenie sodu w podzbiorze wartości typowych wynosi $\bar{x} = 7.70$ mg/dm^3, mediana ma wartość $\tilde{x} = 5.60$ mg/dm^3, a obserwowane wartości mieszczą się w przedziale 0.5–27.2 mg/dm^3. Rozkład stężeń jest logarytmiczno-normalny, współczynnik zmienności ma wartość $V = 84.31\%$.

Wyniki oznaczeń są prawie dwa rzędy wielkości wyższe od granicy oznaczalności DL, zatem cechują się zadowalającą precyzją (tab. 1.3), wariancja techniczna nie przekracza 20% zmienności całkowitej (tab. 1.5), PDL = DL.

Magnez

Stężenia tego wskaźnika w badanym zbiorze zmieniały się w zakresie wartości od 0.29 do 176.90 mg/dm^3. Wartość średnia ($\bar{x}$ = 16.57 mg/dm^3) różni się od 5%-średniej obciętej (14.31 mg/dm^3), co świadczy o wpływie na wartość średnią wartości odstających. Rozkład stężeń magnezu charakteryzuje się asymetrią prawostronną (skośność: 5.329) i spiczastością (kurtoza: 38.749). Współczynnik zmienności ma wartość V = 121.50%.

W zbiorze jest pięć obserwacji ekstremalnych ($x \geqslant$ 44 mg/dm^3; punkty: 11013, 11014, 11019, 11020, 11022). Po ich wyłączeniu z analizy współczynnik zmienności obniżył swą wartość do V = 74.33%.

Rozkład wartości typowych można opisać rozkładem logarytmiczno-normalnym, wartość średnia $\bar{x}$ = 14.18 mg/dm^3 a mediana $\tilde{x}$ = 11.45 mg/dm^3. Obserwowane wartości mieszczą się w przedziale 0.29–42.7 mg/dm^3.

Wyniki oznaczeń magnezu powinny cechować się zadowalającą precyzją, gdyż są dwa rzędy wielkości wyższe od granicy oznaczalności DL (tab. 1.3) a wariancja techniczna nie przekracza 20% zmienności całkowitej (tab. 1.5). Stosunek PDL/DL = 7.

Wapń

Średnie stężenie wapnia w badanych próbkach kształtuje się na poziomie $\bar{x}$ = 79.36 mg/dm^3, przy zmienności tego wskaźnika w granicach od 8.8 do 365.70 mg/dm^3. Rozkład jest asymetryczny (skośność: 2.272, kurtoza: 11.422), w zbiorze są dwie obserwacje ekstremalne (punkty: 11013, 11014). Współczynnik zmienności ma wartość V = 59.35%.

Po wyłączeniu z analizy obserwacji anomalnych współczynnik zmienności maleje do V = 48.29%. Rozkład typowych oznaczeń wapnia jest rozkładem normalnym, z wartością średnią $\bar{x}$ = 76.10 mg/dm^3. Najbardziej typowe wartości (95% przedział ufności dla średniej) mieszczą się w przedziale 70.45–81.75 mg/dm^3.

Wariancja techniczna nie przekracza 20% zmienności całkowitej (tab. 1.5), wyniki oznaczeń są dwa rzędy wielkości wyższe od granicy oznaczalności DL (tab. 1.3), zatem powinny cechować się zadowalającą precyzją.

Azot amonowy

W analizowanym zbiorze stężeń azotu amonowego w próbkach pobranych w ramach RMWP dorzecza górnej Wisły (obszar RZGW Katowice, RZGW Kraków) większość wyników to stężenia poniżej granicy oznaczalności DL. Średnie stężenie tego wskaźnika wynosi $\bar{x}$ = 0.1908 mg/dm^3, przy zmienności w granicach od 0.01 do 3.83 mg/dm^3. Rozkład charakteryzuje się asymetrią prawostronną (skośność: 5.525) i spiczastością (kurtoza: 14.106). Współczynnik zmienności ma wartość V = 184.20%.

16.77% obserwacji (dwadzieścia osiem obserwacji; $x \leqslant$ 0.33 mg/dm^3) to obserwacje ekstremalne (wykres typu „łodyga i liście"). Średnie stężenie analizowanego wskaźnika w podzbiorze wartości anomalnych kształtuje się na poziomie $\bar{x}$ = 0.85 mg/dm^3 przy rozrzucie wartości od 0.33 od 3.83 mg/dm^3.

Średnie stężenie azotu amonowego w podzbiorze wartości typowych $\bar{x}$ = 0.058 mg/dm^3 (mediana: $\tilde{x}$ = 0.04 mg/dm^3). Rozkład obserwacji jest logarytmiczno-normalny, współczynnik zmienności ma wartość V = 105.21%.

Wariancja techniczna obliczona metodą analizy wariancji ANOVA stanowi 7.99% zmienności całkowitej, PDL/DL = 2.5. Ponieważ jednak 21% obserwacji w zbiorze, to wyniki poniżej granicy oznaczalności, a 44% to obserwacje w pobliżu tej granicy, zmienna ta zostanie wyłączona z dalszej analizy.

Glin

Średnie stężenie glinu kształtuje się na poziomie $\bar{x} = 0.059$ mg/dm^3, przy rozrzucie wartości od 0.006 do 0.470 mg/dm^3. Rozkład stężeń glinu charakteryzuje się silną asymetrią prawostronną (skośność: 3.532) i dużą spiczastością (kurtoza: 21.375). Współczynnik zmienności ma wartość $V = 89.71\%$.

Cztery obserwacje z badanego zbioru (punkty: 21036, 21067, 21078, 21080) to obserwacje odstające ($x \geqslant 0.184$ mg/dm^3). Po ich wyłączeniu z analizy współczynnik zmienności obniżył swą wartość $V = 67.19\%$. Rozkład typowych stężeń glinu w badanych próbkach jest rozkładem logarytmiczno-normalnym z wartością średnią $\bar{x} = 0.054$ mg/dm^3 i medianą $\tilde{x} = 0.047$ mg/dm^3.

Wariancja techniczna wyznaczona z wykorzystaniem analizy wariancji ANOVA stanowi 27% zmienności całkowitej (tab. 1.5). Wyniki są tego samego rzędu wielkości co deklarowana przez laboratoria granica oznaczalności DL, stosunek PDL/DL = 5, zatem zmienna ta zostanie wyłączona z dalszej analizy.

Żelazo ogólne

Stężenia tego wskaźnika w analizowanych próbkach w większości przypadków są stężeniami mniejszymi od granicy oznaczalności DL lub wynikami w pobliżu tej granicy (52% obserwacji). Wartość średnia wynosi $\bar{x} = 1.62$ mg/dm^3 przy rozrzucie wartości od 0.005 do 48.50 mg/dm^3. Rozkład cechuje się silną asymetrią prawostronną (skośność: 6.995) i znaczną spiczastością (kurtoza: 64.625). Współczynnik zmienności wynosi $V = 284.99\%$.

20.36% obserwacji (trzydzieści cztery obserwacje; $x \geqslant 2.4$ mg/dm^3) to obserwacje ekstremalne (wykres „łodyga i liście" oraz „skrzynka z wąsami"). Średnie stężenie żelaza w podzbiorze obserwacji anomalnych kształtuje się na poziomie $\bar{x} = 7.22$ mg/dm^3, mediana ma wartość $\tilde{x} = 4.99$ mg/dm^3. Podzbiór ten charakteryzuje się zakresem wartości od 2.40 do 48.5 mg/dm^3.

Średnie typowe stężenie żelaza ogólnego wynosi $\bar{x} = 0.19$ mg/dm^3, a współczynnik zmienności ma wartość $V = 181.60\%$.

Wariancja techniczna obliczona metodą klasyczną (z wykorzystaniem analizy wariancji) stanowi 45% zmienności całkowitej, co oznacza niezadowalającą precyzję oznaczeń. Wyniki oznaczeń są tego samego rzędu wielkości co granica oznaczalności, stosunek PDL/DL = 73. Zmienna ta zostanie wyłączona z dalszej analizy.

Mangan

Średnie stężenie manganu w badanym zbiorze wynosi $\bar{x} = 0.315$ mg/dm^3, przy zmienności w zakresie od 0.005 do 26.400 mg/dm^3. Rozkład jest asymetryczny prawostronnie (skośność: 12.428) i bardzo spiczasty (kurtoza: 158.252), współczynnik zmienności $V = 653.91\%$.

W zbiorze są dwadzieścia dwie obserwacje odstające ($x \geqslant 0.41$ mg/dm^3), co stanowi 13.17% wszystkich obserwacji. Podzbiór obserwacji anomalnych charakteryzuje się wartością średnią $\bar{x} = 2.12$ mg/dm^3, medianą $\tilde{x} = 0.76$ mg/dm^3 i rozrzutem wartości od 0.41 do 26.4 mg/dm^3.

Rozkład typowych wartości stężeń manganu nie jest rozkładem normalnym ani logarytmiczno-normalnym, wartość średnia wynosi $\bar{x} = 0.05$ mg/dm^3 a mediana $\tilde{x} = 0.01$ mg/dm^3. Współczynnik zmienności ma wartość $V = 162.58\%$.

Wariancja techniczna obliczona metodą klasyczną (analizy wariancji) stanowi aż 70% zmienności całkowitej (tab. 1.5). Stosunek PDL/DL = 3.

56% stanowią wyniki w pobliżu lub poniżej deklarowanej przez laboratoria granicy oznaczalności (tab. 1.3), zmienna ta zostanie wyłączona z dalszej analizy.

Azot azotynowy

Stężenia azotu azotynowego zmieniają się w zakresie 0.001–0.084 mg/dm^3, zaś średnie stężenie kształtuje się na poziomie $\bar{x} = 0.004$ mg/dm^3. Rozkład charakteryzuje się silną asymetrią prawostronną (skośność: 8.437) i spiczastością (kurtoza: 89.420). Współczynnik zmienności badanego zbioru ma wartość $V = 181.61\%$.

Czternaście obserwacji z analizowanego zbioru (8.38% wszystkich obserwacji) to obserwacje ekstremalne ($x \geqslant 0.009$ mg/dm^3). Tworzą one podzbiór wartości anomalnych, charakteryzujący się wartością średnią $\bar{x} = 0.019$ mg/dm^3, medianą $\tilde{x} = 0.015$ mg/dm^3 i rozrzutem wartości 0.009–0.084 mg/dm^3.

Podzbiór wartości typowych można opisać rozkładem logarytmiczno-normalnym o wartości średniej $\bar{x} = 0.0026$ mg/dm^3. Mediana ma wartość $\tilde{x} = 0.002$ mg/dm^3, a współczynnik zmienności $V = 67.87\%$.

Wyniki oznaczeń azotu azotynowego nie będą cechowały się zadowalającą precyzją — są to wyniki tego samego rzędu wielkości co granica oznaczalności DL, bądź wyniki poniżej tej granicy (tab. 1.3). Wariancja techniczna σ^2_{tech} obliczona metodą analizy wariancji ANOVA stanowi wprawdzie 3% zmienności całkowitej (tab. 1.5), stosunek PDL/DL = 3, jednak zmienna ta zostanie wyłączona z dalszej analizy (34% wyników stanowią tu oznaczenia poniżej granicy oznaczalności DL).

Azot azotanowy

Średnie stężenie azotu azotanowego w analizowanych próbkach kształtuje się na poziomie $\bar{x} = 2.75$ mg/dm^3, przy rozrzucie wartości od 0.1 do 16.5 mg/dm^3 (współczynnik zmienności wynosi $V = 127.82\%$). Rozkład charakteryzuje się asymetrią prawostronną (skośność: 1.817) i spiczastością (kurtoza: 3.230).

Z wykresu typu „łodyga i liście" wynika, że w badanym zbiorze jest siedem obserwacji ekstremalnych ($x \geqslant 11.2$ mg/dm^3; punkty: 11010, 11030, 11046, 21016, 21023, 21034, 21072).

Po wyłączeniu z analizy obserwacji anomalnych współczynnik zmienności obniżył swą wartość do $V = 116.76\%$. Średnie typowe stężenie azotu azotanowego w badanych próbkach kształtuje się na poziomie $\bar{x} = 2.26$ mg/dm^3, przy rozrzucie wartości w przedziale od 0.1 do 9.5 mg/dm^3 i wartości środkowej $\tilde{x} = 1.2$ mg/dm^3.

W zbiorze tym 34% obserwacji to wyniki poniżej granicy oznaczalności DL deklarowanej przez laboratoria, pozostałe wyniki są jeden rząd wielkości wyższe od granicy oznaczalności (tab. 1.3). Wariancja techniczna σ^2_{tech} stanowi wprawdzie 7% wariancji całkowitej (tab. 1.5), jednak zmienna ta zostanie wyłączona z dalszej analizy.

Chlorki

Stężenia chlorków w analizowanych próbkach mieszczą się w zakresie 3.00–372.00 mg/dm^3, przy wartości średniej $\bar{x} = 27.73$ mg/dm^3 i współczynniku zmienności $V = 147.56\%$. 5%-średnia obcięta ma wartość 21.96 mg/dm^3, co oznacza że na średnią mają wpływ obserwacje odstające. Rozkład badanych stężeń jest asymetryczny (skośność: 5.661) i spiczasty (kurtoza: 40.386).

Z wykresu łodyga i liście wynika, że w zbiorze wyników stężeń chlorków w próbkach pobranych w ramach RMWP w obszarach RZGW Katowice i RZGW Kraków jest dziesięć obserwacji ekstremalnych ($x \geqslant 78$ mg/dm^3).

Po wyłączeniu z analizy obserwacji anomalnych współczynnik zmienności obniżył swą wartość $V = 74.73\%$. Typowe stężenia chlorków w analizowanych próbkach mają w przybliżeniu rozkład logarytmiczno-normalny, z wartością średnią $\bar{x} = 20.4$ mg/dm^3, medianą $\tilde{x} = 16.3$ mg/dm^3 i zakresem zmienności 3.0–72.2 mg/dm^3.

Wyniki oznaczeń chlorków w badanych próbkach mają ten sam rząd wielkości co deklarowane przez laboratoria granice oznaczalności (tab. 1.3), wartość PDL = DL, a wariancja techniczna σ^2_{tech} stanowi 15% zmienności całkowitej (tab. 1.5).

Siarczany

Średnie stężenie siarczanów w analizowanym zbiorze wynosi $\bar{x} = 61.72$ mg/dm^3 przy rozrzucie wartości 10.00–1235.00 mg/dm^3. 5%-średnia obcięta ma wartość 44.90 mg/dm^3, co oznacza wpływ wartości ekstremalnych na wartość średnią. Rozkład charakteryzuje się silną asymetrią prawostronną (skośność: 7.734) i spiczastością (kurtoza: 65.194), współczynnik zmienności ma wartość $V = 209.47\%$.

Z wykresu typu „łodyga i liście" wynika, że w zbiorze jest siedem obserwacji ekstremalnych (obserwacje w przedziale 149.00–1235.00 mg/dm^3). Po ich wyłączeniu z analizy współczynnik zmienności obniżył swą wartość do $V = 75.85\%$. Średnie stężenie siarczanów kształtuje się na poziomie $\bar{x} = 43.85$ mg/dm^3 przy rozrzucie wartości 10.00–136.00 mg/dm^3 (mediana: $\tilde{x} = 29.1$ mg/dm^3).

Wyniki oznaczeń siarczanów są tego samego rzędu wielkości co DL lub o rząd wielkości wyższe od deklarowanej przez laboratoria granicy oznaczalności DL (tab. 1.3), PDL = DL, a wariancja techniczna stanowi 2% zmienności całkowitej (tab. 1.5). Oznacza to, że wyniki te będą cechować się zadowalającą precyzją.

Fosforany rozpuszczone

Stężenia fosforanów mieszczą się w zakresie od 0.05 do 1.40 mg/dm^3, średnie stężenie wynosi $\bar{x} = 0.16$ mg/dm^3, współczynnik zmienności ma wartość $V = 139.74\%$. Rozkład wyników jest skośny prawostronnie (skośność: 3.105) i spiczasty (kurtoza: 11.252).

15.6% obserwacji w zbiorze (dwadzieścia sześć obserwacji; $x \geqslant 0.31$ mg/dm^3) to obserwacje ekstremalne. Stanowią one podzbiór obserwacji anomalnych o wartości średniej $\bar{x} = 0.58$ mg/dm^3, medianie $\tilde{x} = 0.49$ mg/dm^3 i rozrzucie wartości 0.31–1.40 mg/dm^3.

Po wyłączeniu z analizy obserwacji anomalnych współczynnik zmienności obniżył swą wartość do $V = 64.84\%$. Średnie stężenie fosforanów rozpuszczonych w podzbiorze obserwacji typowych kształtuje się na poziomie $\bar{x} = 0.078$ mg/dm^3 przy rozrzucie wartości od 0.05 do 0.27 mg/dm^3.

Wariancja techniczna σ^2_{tech} stanowi zaledwie kilka procent zmienności całkowitej (tab. 1.5), PDL = DL, jednak prawie 90% oznaczeń stanowią wyniki poniżej deklarowanej przez laboratoria granicy oznaczalności DL (tab. 1.3), zatem zmienna ta zostanie wyłączona z analizy.

Krzemionka zdysocjowana

Stężenia tego wskaźnika mieszczą się w zakresie 0.50–31.60 mg/dm^3, a wartość średnia kształtuje na poziomie $\bar{x} = 12.44$ mg/dm^3. Rozkład jest w przybliżeniu normalny (skośność: 0.738, kurtoza: -0.414), współczynnik zmienności ma wartość $V = 65.97\%$.

Na wykresach (wykres „łodyga i liście", „skrzynka z wąsami") nie ma obserwacji ekstremalnych.

Wyniki oznaczeń krzemionki zdysocjowanej charakteryzują się zadowalającą precyzją, gdyż są rząd wielkości wyższe od granicy oznaczalności DL deklarowanej przez laboratoria (tab. 1.3), wartość PDL = DL. Wariancja techniczna, wyznaczona metodą analizy wariancji ANOVA stanowi zaledwie 4% zmienności całkowitej (tab. 1.5).

Fluorki

Średnie stężenie tego wskaźnika w analizowanych próbkach kształtuje się na poziomie $\bar{x} = 0.199$ mg/dm^3, przy rozrzucie wartości od 0.01 do 0.81 mg/dm^3 (współczynnik zmienności $V = 80.75\%$). Rozkład charakteryzuje się asymetrią prawostronną (skośność: 1.151) i spiczastością (kurtoza: 1.520).

Z wykresu typu „łodyga i liście" wynika, że w zbiorze są cztery wartości ekstremalne ($x \geqslant 0.62$ mg/dm^3; punkty: 21008, 21058, 21075, 21094).

Po wyłączeniu z analizy obserwacji ekstremalnych współczynnik zmienności nieznacznie obniżył swą wartość $V = 74.72\%$. Średnie stężenie fluorków w podzbiorze wartości typowych kształtuje się na poziomie $\bar{x} = 0.186$ mg/dm^3 przy rozrzucie wartości 0.01–0.54 mg/dm^3 (mediana: $\tilde{x} = 0.17$ mg/dm^3).

W przypadku tej zmiennej porównywalność badań jest utrudniona, gdyż laboratoria deklarowały o rząd wielkości różniące się granice oznaczalności DL (tab. 1.3). Wyniki poniżej granic oznaczalności stanowią ok. 10% wszystkich wyników oznaczeń, uzyskane wyniki są w większości przypadków jeden rząd wielkości wyższe od granic oznaczalności, wartość PDL = DL, a wariancja techniczna σ^2_{tech} stanowi 3% zmienności całkowitej (tab. 1.5).

Bor

Wskaźnik ten nie był w ogóle oznaczany w próbkach pobranych w obszarze RZGW Katowice, a spośród punktów z obszaru RZGW Kraków był oznaczony tylko w 41 przypadkach, z czego ok. 80% to wyniki poniżej granicy oznaczalności DL — zostanie zatem przeprowadzona analiza statystyczna istniejącego zbioru danych, ale zmienna ta będzie wyłączona z dalszej analizy.

Wartości oznaczeń boru mieszczą się w zakresie od 0.005 do 0.825 mg/dm^3, przy wartości średniej $\bar{x} = 0.077$ mg/dm^3 i bardzo dużym współczynniku zmienności $V = 201.46\%$. Rozkład wartości stężeń boru charakteryzuje się asymetrią prawostronną (skośność: 3.394), i znaczną spiczastością (kurtoza: 13.320).

Z wykresu typu „łodyga i liście" wynika, że w analizowanym zbiorze jest siedem obserwacji ekstremalnych ($x \geqslant 0.168$ mg/dm^3). Po ich wyłączeniu z analizy współczynnik zmienności dla zbioru ma wartość $V = 135.91\%$. Średnie stężenie boru w podzbiorze wartości typowych kształtuje się na niższym poziomie $\bar{x} = 0.022$ mg/dm^3, przy rozrzucie wartości 0.005–0.130 mg/dm^3.

W przypadku tej zmiennej, ze względu na małą ($N < 11$) liczbę par wyników próbek dublowanych, nie można było dokonać oceny precyzji oznaczeń z wykorzystaniem programu ROB 2.

Chrom ogólny

W przypadku tej zmiennej w 153 przypadkach na 166 obserwacji (jeden brak danych), wyniki oznaczeń są poniżej granicy oznaczalności DL — zmienna ta nie będzie uwzględniana w dalszej analizie.

Stężenia chromu w analizowanym zbiorze zmieniają się w zakresie 0.001–0.010 mg/dm^3, przy wartości średniej $\bar{x} = 0.0041$ mg/dm^3 (współczynnik zmienności $V = 102.92\%$).

Według wykresu typu „łodyga i liście" w zbiorze nie obserwacji ekstremalnych. Wariancja techniczna σ^2_{tech} stanowi ponad 20% zmienności całkowitej.

Cynk

Średnie stężenie cynku w badanych próbkach kształtuje się na poziomie $\bar{x} = 0.231$ mg/dm^3, przy rozrzucie wartości 0.003–15.000 mg/dm^3. Rozkład jest silnie asymetryczny (skośność: 10.174) i spiczasty (kurtoza: 110.748), współczynnik zmienności ma bardzo dużą wartość $V = 555.97\%$.

12% obserwacji (dwadzieścia obserwacji; $x \geqslant 0.22$ mg/dm^3) to obserwacje ekstremalne (wykres typu „łodyga i liście"). Tworzą one podzbiór obserwacji anomalnych, w którym wartość średnia $\bar{x} = 0.147$ mg/dm^3 a mediana wynosi $\tilde{x} = 0.363$ mg/dm^3.

Podzbiór obserwacji typowych charakteryzuje się logarytmiczno-normalnym rozkładem, o wartości średniej $\bar{x} = 0.049$ mg/dm^3, medianie $\tilde{x} = 0.033$ mg/dm^3 i współczynniku zmienności $V = 88.76\%$.

Wyniki oznaczeń cynku charakteryzują się niską precyzją oznaczeń, są to dane tego samego rzędu wielkości co granica oznaczalności deklarowana przez laboratoria DL (tab. 1.3).

Wariancja techniczna σ^2_{tech} — wyznaczona z wykorzystaniem analizy wariancji ANOVA — stanowi 87% zmienności całkowitej (tab. 1.5).

Cynk jest wskaźnikiem, który z uwagi na łatwość migracji powszechnie występuje w wodach podziemnych. Nie będzie on wyłączony z dalszej analizy, co pozwoli sprawdzić jakość prognoz uzyskiwanych dla wskaźników fizyko-chemicznych wód charakteryzujących się niską precyzją oznaczeń.

Kadm

W 143 przypadkach na 164 obserwacje (trzy braki danych) są to wyniki poniżej granicy oznaczalności DL (87% oznaczeń). Stężenia kadmu zmieniają się w zakresie 0.001–0.056 mg/dm^3. Wartość średnia: $\bar{x} = 0.00187$ mg/dm^3, a współczynnik zmienności: $V = 248.72\%$. Rozkład charakteryzuje się silną asymetrią prawostronną (skośność: 10.072), i znaczną spiczastością (kurtoza: 115.142).

Wszystkie obserwacje powyżej granicy oznaczalności są potraktowane jako obserwacje odstające ($x \geqslant 0.001$ mg/dm^3). Zmienna ta nie będzie uwzględniona w dalszej analizie.

Miedź

Średnie stężenie miedzi w badanych próbkach kształtuje się na poziomie $\bar{x} = 0.0061$ mg/dm^3, przy rozrzucie wartości 0.001–0.040 mg/dm^3 (współczynnik zmienności $V = 74.36\%$). Rozkład jest prawostronnie asymetryczny (skośność: 3.417) i spiczasty (kurtoza: 21.393).

Dwie obserwacje spośród 165 (dwa braki danych) zakwalifikowane zostały jako obserwacje ekstremalne ($x \geqslant 0.03$ mg/dm^3; punkty: 11013, 11014). Po ich wyłączeniu z analizy współczynnik zmienności ma wartość $V = 55.56\%$, przy wartości średniej $\bar{x} = 0.0058$ mg/dm^3 i medianie $\tilde{x} = 0.0050$ mg/dm^3.

W przypadku tej zmiennej ocena precyzji jest utrudniona ze względu na różne granice oznaczalności DL deklarowane przez laboratoria (różniące się o rząd wielkości). Ponieważ we wszystkich próbkach pobieranych w obszarze RZGW Katowice (33% obserwacji) stężenie miedzi było poniżej granicy oznaczalności DL (tab. 1.3), zmienna ta zostanie wyłączona z dalszej analizy.

W obszarze RZGW Kraków PDL = 5DL, a wyniki są tego samego rzędu wielkości co laboratoryjna granica oznaczalności DL. Wariancja techniczna σ^2_{tech} — wyznaczona z wykorzystaniem analizy wariancji ANOVA — stanowi 25% wariancji całkowitej (tab. 1.5).

Nikiel

Wyniki oznaczeń niklu w analizowanych próbkach mieszczą się w zakresie od 0.001 do 0.071 mg/dm^3. Średnie stężenie tego wskaźnika w badanym zbiorze kształtuje się na poziomie $\bar{x} = 0.0064$ mg/dm^3 (współczynnik zmienności $V = 133.76\%$). Rozkład charakteryzuje się asymetrią prawostronną (skośność: 4.931) i spiczastością (kurtoza: 31.050).

Na wykresie typu „łodyga i liście" widoczne są trzy obserwacje ekstremalne (wartości $x \geqslant 0.05$ mg/dm^3; punkty: 11013, 21067, 21081). Po ich wyłączeniu z analizy współczynnik zmienności obniżył swą wartość do $V = 79.81\%$. Średnie stężenie niklu kształtuje się na poziomie $\bar{x} = 0.0054$ mg/dm^3 (mediana: $\tilde{x} = 0.0040$ mg/dm^3).

Laboratoria wykonujące analizy w ramach RMWP dorzecza górnej Wisły deklarowały różne granice oznaczalności tego wskaźnika — różniące się o rząd wielkości. Ponieważ w próbkach z obszaru RZGW Katowice uzyskano, poza dwoma wyjątkami, wyniki poniżej granicy oznaczalności DL (32% obserwacji; tab. 1.3), zatem zmienna ta nie będzie uwzględniana w dalszej analizie.

Ołów

Stężenia ołowiu mieszczą się w przedziale od 0.001 do 0.030 mg/dm^3, z wartością średnią $\bar{x} = 0.00549$ mg/dm^3 (współczynnik zmienności ma wartość $V = 89.61\%$). Rozkład cechuje asymetria prawostronna (skośność: 1.715) i spiczastość (kurtoza: 4.561).

Z wykresu typu „łodyga i liście" wynika, że w zbiorze są dwie obserwacje ekstremalne (punkty: 11044, 21100). Po ich wyłączeniu z analizy współczynnik zmienności ma wartość $V = 82.07\%$, a średnie stężenie ołowiu kształtuje się na poziomie $\bar{x} = 0.0052$ mg/dm^3.

Wyniki oznaczeń ołowiu są tego samego rzędu wielkości co laboratoryjna granica oznaczalności DL, stosunek PDL/DL ≈ 7, a wariancja techniczna stanowi 70% zmienności całkowitej (tab. 1.5). Wiadmo również, że próbki były zanieczyszczane ołowiem w trakcie transportu (Witczak et al., 1994), zatem zmienna ta zostanie wyłączona z dalszej analizy.

Rtęć

Średnie stężenie rtęci w badanych próbkach kształtuje się na poziomie $\bar{x} = 0.00119$ mg/dm^3, przy rozrzucie wartości od 0.0002 do 0.0056 mg/dm^3 (współczynnik zmienności ma wartość $V = 106.94\%$). Rozkład jest skośny (skośność: 1.317) i spiczasty (kurtoza: 1.255).

W zbiorze są trzy obserwacje ekstremalne (wartości $x \geqslant 0.0055$ mg/dm^3), po ich wyłączeniu z analizy współczynnik zmienności nieznacznie obniżył swoją wartość $V = 102.66\%$, średnie stężenie rtęci wynosi $\bar{x} = 0.0011$ mg/dm^3, a mediana ma wartość $\tilde{x} = 0.0005$ mg/dm^3.

Ze względu na niską precyzję oznaczeń — 45% oznaczeń to wyniki poniżej granicy oznaczalności DL (tab. 1.3), stosunek PDL/DL = 19, a wariancja techniczna σ^2_{tech} stanowi 27% wariancji całkowitej (tab. 1.5), zmienna ta zostanie wyłączona z dalszej analizy.

Dwutlenek węgla agresywny

W badanych próbkach stężenia dwutlenku węgla agresywnego (wolnego dwutlenku węgla rozpuszczonego w wodzie, nadmiarowego w stosunku do dwutlenku węgla biernego, potrzebnego do utrzymania w stanie rozpuszczonym, zdysocjowanym, wodorowęglanów — Witczak, Adamczyk, 1994) mieściły się w zakresie 2.20–41.40 mg/dm^3, przy wartości średniej $\bar{x} = 6.627$ mg/dm^3 (współczynnik zmienności $V = 110.94\%$). Rozkład charakteryzuje się asymetrią prawostronną (skośność: 2.780) i spiczastością (kurtoza: 8.570).

Na wykresie typu „łodyga i liście" widoczne jest jedenaście obserwacji ekstremalnych (wartości $x \geqslant 17.2$ mg/dm^3). Stanowią one podzbiór obserwacji anomalnych, charakteryzujący się wartością średnią $\bar{x} = 26.84$ mg/dm^3 i medianą $\tilde{x} = 23.10$ mg/dm^3.

Wartość średnia w podzbiorze obserwacji typowych wynosi $\bar{x} = 4.76$ mg/dm^3, mediana ma wartość $\tilde{x} = 10.00$ mg/dm^3 a współczynnik zmienności $V = 68.91\%$.

Wariancja techniczna obliczona metodą analizy wariancji stanowi 28% zmienności całkowitej (tab. 1.5), w obszarze RZGW Kraków występuje duża liczba braków danych (22% obserwacji), zatem zmienna ta zostanie wyłączona z dalszej analizy.

Współczynnik absorpcji UV (A 254)

Wartość 5%-średniej obciętej $\bar{x} = 0.11124$ odbiega od wartości średniej $\bar{x} = 0.15089$, co świadczy o wpływie na wartość średnią obserwacji ekstremalnych. Współczynnik absorpcji UV zmienia się w zakresie od 0.001 do 3.000. Rozkład cechuje się asymetrią prawostronną (skośność: 6.557) i spiczastością (kurtoza: 58.737). Współczynnik zmienności ma dużą wartość $V = 191.17\%$.

W badanym zbiorze jest 10.18% obserwacji ekstremalnych (osiemnaście obserwacji; wartości $x \geqslant 0.38$), stanowią one podzbiór obserwacji anomalnych o wartości średniej $\bar{x} = 0.70$ i medianie $\tilde{x} = 0.52$.

Po wyłączeniu z analizy obserwacji odstających (podzbiór obserwacji typowych) średnia wartość współczynnika absorpcji UV wynosi $\bar{x} = 0.0084$, mediana ma wartość $\tilde{x} = 0.052$, a współczynnik zmienności $V = 105.87\%$.

Zmienna ta charakteryzuje się zadowalającą precyzją oznaczeń (wariancja techniczna σ^2_{tech} stanowi mniej niż 20% zmienności całkowitej — tab. 1.5), uzyskane wyniki są o rząd wielkości wyższe od granicy oznaczalności DL (tab. 1.3), stosunek PDL/DL ≈ 6.

Rozpuszczony węgiel organiczny

Wartości stężeń rozpuszczonego węgla organicznego mieszczą się w zakresie od 0.20 do 14.00 mg/dm^3, przy wartości średniej $\bar{x} = 1.67$ mg/dm^3 (5%-średnia obcięta ma wartość 1.43 mg/dm^3 co oznacza wpływ na wartość średnią obserwacji ekstremalnych). Rozkład badanych stężeń jest asymetryczny prawostronnie (skośność: 4.604) i spiczasty (kurtoza: 28.396). Współczynnik zmienności wynosi $V = 103.46\%$.

Aż jedenaście obserwacji to obserwacje ekstremalne ($x \geqslant 3.7$ mg/dm^3), stanowią one podzbiór obserwacji anomalnych, w którym wartość średnia $\bar{x} = 6.38$ mg/dm^3 a mediana $\tilde{x} = 4.71$ mg/dm^3.

Podzbiór obserwacji typowych (po wyłączeniu obserwacji ekstremalnych) charakteryzuje się rozkładem logarytmiczno-normalnym z wartością średnią $\bar{x} = 1.33$ mg/dm^3, medianą $\tilde{x} = 1.12$ mg/dm^3 i współczynnikiem zmienności $V = 56.49\%$.

Wariancja techniczna σ^2_{tech} nie przekracza 20% zmienności całkowitej (tab. 1.5), wartości wyników są tego samego rzędu wielkości co granica oznaczalności DL deklarowana przez laboratoria (tab. 1.3), PDL/DL ≈ 12.

Utlenialność ChZT

Parametr ten zmienia się w zakresie 0.5–17.0 mg/dm^3, wartość średnia $\bar{x} = 1.85$ mg/dm^3 a współczynnik zmienności $V = 97.08\%$. Rozkład charakteryzuje się asymetrią prawostronną (skośność: 4.700) i spiczastością (kurtoza: 32.467).

Z wykresu typu „łodyga i liście" wynika, że czternaście obserwacji, to obserwacje odstające (wartości $x \geqslant 3.9$ mg/dm^3). Stanowią one subpopulację anomalną o wartości średniej $\bar{x} = 6.16$ mg/dm^3 i medianie $\tilde{x} = 4.90$ mg/dm^3.

Subpopulację typową można opisać w przybliżeniu rozkładem logarytmiczno-normalnym, z wartością średnią $\bar{x} = 1.45$ mg/dm^3, medianą $\tilde{x} = 1.20$ mg/dm^3 i współczynnikiem zmienności $V = 52.48\%$.

Wariancja techniczna σ^2_{tech} stanowi kilka procent wariancji całkowitej (tab. 1.5). Wyniki oznaczeń mają ten sam rząd wielkości co granica oznaczalności DL deklarowana przez laboratoria (tab. 1.3), stosunek PDL/DL = 2.

Azot organiczny

Wartość średnia tego parametru to $\bar{x} = 0.393$ mg/dm^3, przy rozrzucie wartości od 0.01 do 4.12 mg/dm^3 (współczynnik zmienności $V = 134.36\%$). Rozkład jest asymetryczny prawostronnie (skośność: 3.813) i spiczasty (kurtoza: 19.522).

W badanym zbiorze jest dziesięć obserwacji ekstremalnych (wartości $x \geqslant 1.12$ mg/dm^3), które tworzą podzbiór wartości anomalnych o wartości średniej $\bar{x} = 2.03$ mg/dm^3 i medianie $\tilde{x} = 0.006$ mg/dm^3.

Podzbiór wartości typowych można opisać rozkładem logarytmiczno-normalnym. Średnie stężenie azotu organicznego w badanych próbkach wynosi $\bar{x} = 0.288$ mg/dm^3, mediana ma wartość $\tilde{x} = 0.21$ mg/dm^3, a współczynnik zmienności $V = 84.63\%$.

Wyniki reprezentują ten sam rząd wielkości co granica oznaczalności DL (tab. 1.3), wartość PDL = DL, jednak wariancja techniczna σ^2_{tech} obliczona metodą analizy wariancji ANOVA stanowi 58% wariancji całkowitej (tab. 1.5), zmienna ta zostanie wyłączona z dalszej analizy.

Fenole lotne

W zdecydowanej większości przypadków (77% obserwacji, 128 na 166 obserwacji, 1 brak danych) stężenia fenoli lotnych w badanych próbkach to stężenia poniżej granicy oznaczalności DL — zmienna ta nie będzie zatem uwzględniana w dalszej analizie.

Współczynnik zmienności badanego zbioru wynosi $V = 228.39\%$. Na wykresie typu „łodyga i liście" wszystkie próbki (próbki z obszaru RZGW Katowice) w których stężenia fenoli kształtowały się powyżej granicy oznaczalności zostały zakwalifikowane jako obserwacje ekstremalne. Wariancja techniczna stanowi aż 64% wariancji całkowitej (tab. 1.5), PDL = DL.

Substancje ropopochodne

Średnie stężenie substancji ropopochodnych kształtowało się na poziomie $\bar{x} = 0.33$ mg/dm^3, przy rozrzucie wartości 0.01–2.38 mg/dm^3 (współczynnik zmienności $V = 138.39\%$). Rozkład charakteryzuje się asymetrią prawostronną (skośność: 2.541) i spiczastością (kurtoza: 6.999).

W badanym zbiorze jest trzynaście obserwacji odstających (wartości $x \geqslant 1.06$ mg/dm^3). Tworzą one podzbiór obserwacji anomalnych o wartości średniej $\bar{x} = 1.62$ mg/dm^3 i medianie $\tilde{x} = 1.42$ mg/dm^3.

Subpopulację typową można opisać rozkładem logarytmiczno-normalnym o wartości średniej $\bar{x} = 0.22$ mg/dm^3 i medianie $\tilde{x} = 0.12$ mg/dm^3 ($V = 102.50\%$).

Uzyskane wyniki oznaczeń mają ten sam rząd wielkości co granica oznaczalności (tab. 1.3), PDL = 9DL, wariancja techniczna nie przekracza 20% wariancji całkowitej (tab. 1.5). Z dokumentacji wynika, że na skutek przebywania w trakcie transportu w atmosferze spalin rejestrowano podwyższone stężenia tego wskaźnika (Witczak et al., 1994) zmienna ta zostanie zatem wyłączona z dalszej analizy.

Substancje powierzchniowo-czynne anionowe

W obszarze RZGW Kraków we wszystkich analizowanych próbkach (33% obserwacji) stężenie substancji powierzchniowo-czynnych anionowych kształtowało się na poziomie poniżej granicy oznaczalności DL = 0.1 mg/dm^3, zatem zmienna ta nie będzie uwzględniana w dalszej analizie, zostanie przeprowadzona tylko ocena statystyczna zbioru.

Średnie stężenie substancji powierzchniowo-czynnych anionowych kształtuje się na poziomie $\bar{x} = 0.078$ mg/dm^3, przy rozrzucie wartości 0.01–0.12 mg/dm^3. Współczynnik zmienności wynosi $V = 43.39\%$.

Chloroform

Stężenie chloroformu w analizowanych próbkach kształtowało się na poziomie od 0.00001 do 1.1700 mg/dm^3, przy wartości średniej $\bar{x} = 0.069$ mg/dm^3 (współczynnik zmienności $V = 236.40\%$). Rozkład wartości charakteryzuje się asymetrią prawostronną (skośność: 3.726) i spiczastością (kurtoza: 17.432).

Z analizy wykresu typu „łodyga i liście" wynika, że 15.6% obserwacji (dwadzieścia sześć obserwacji) należy zakwalifikować jako ekstremalne (wartości $x \geqslant 0.080$ mg/dm^3). Tworzą one podzbiór obserwacji anomalnych, w którym $\bar{x} = 0.38$ mg/dm^3 i $\tilde{x} = 0.34$ mg/dm^3.

Współczynnik zmienności dla podzbioru wartości typowych (po wyłączeniu z analizy obserwacji ekstremalnych) ma wartość $V = 128.31\%$, rozkład jest w przybliżeniu rozkładem logarytmiczno-normalnym. Średnie stężenie chloroformu w tych próbkach kształtuje się na poziomie $\bar{x} = 0.011$ mg/dm^3 (mediana $\tilde{x} = 0.005$ mg/dm^3).

Wariancja techniczna σ^2_{tech} obliczona metodą klasycznej analizy wariancji stanowi 22% wariancji całkowitej. Uzyskane wyniki są trzy rzędy wielkości wyższe od granicy oznaczalności DL. Z kolei praktyczna granica oznaczalności PDL jest cztery rzędy wielkości wyższa od granicy oznaczalności DL. Jak wynika z dokumentacji (Witczak et al., 1994), we wszystkich próbkach (nawet próbkach zerowych) obserwowano podwyższone stężenia tego wskaźnika, zatem zmienna ta zostanie wyłączona z dalszej analizy.

Tetrachloroetylen

Średnie stężenie tego wskaźnika w analizowanych próbkach wynosi $\bar{x} = 0.00035$ mg/dm^3, przy rozrzucie wartości od 0.000005 do 0.0067 mg/dm^3 (współczynnik zmienności ma wartość $V = 223.81\%$). Rozkład jest asymetryczny (skośność: 5.122) i spiczasty (kurtoza: 32.809).

W badanym zbiorze stężeń tetrachloroetylenu 12.6% obserwacji (dwadzieścia jeden obserwacji), to wartości ekstremalne (wartości $x \geqslant 0.00081$ mg/dm^3). Tworzą one podzbiór obserwacji anomalnych o wartości średniej $\bar{x} = 0.0018$ mg/dm^3 i medianie $\tilde{x} = 0.0012$ mg/dm^3.

Wartość średnia dla podzbioru obserwacji typowych wynosi $\bar{x} = 0.00014$ mg/dm^3, mediana ma wartość $\tilde{x} = 0.00006$ mg/dm^3 a współczynnik zmienności $V = 118.47\%$.

Wyniki poniżej granicy oznaczalności stanowią 37% obserwacji. Ze względu na niską precyzję oznaczeń (tab. 1.3, 1.5) zmienna ta nie będzie uwzględniona w dalszej analizie.

Trójchloroetylen

Zbiór wartości stężeń trójchloroetylenu w badanych próbkach zawiera się w przedziale od 0.00003 do 0.08820 mg/dm^3, przy średnim stężeniu na poziomie $\bar{x} = 0.00145$ mg/dm^3 i bardzo dużym współczynniku zmienności $V = 589.03\%$. Rozkład wartości charakteryzuje się silną asymetrią prawostronną (skośność: 8.758) i znaczną spiczastością (kurtoza: 80.770).

Z wykresu typu „łodyga i liście" wynika, że trzydzieści obserwacji (18% wartości) to obserwacje ekstremalne ($x \geqslant 0.00034$ mg/dm^3). Stanowią one podzbiór obserwacji anomalnych o wartości średniej $\bar{x} = 0.0077$ mg/dm^3 i medianie $\tilde{x} = 0.0019$ mg/dm^3.

Podzbiór wartości typowych charakteryzuje się wartością średnią $\bar{x} = 0.000063$ mg/dm^3 i medianą $\tilde{x} = 0.00003$ mg/dm^3. Współczynnik zmienności $V = 92.14\%$.

Dużą grupę obserwacji (49% obserwacji) stanowią wyniki poniżej granicy oznaczalności DL, praktyczna granica oznaczalności PDL jest trzy rzędy wielkości wyższa do DL (tab. 1.3), wariancja techniczna σ^2_{tech} obliczona metodą klasyczną (analizy wariancji) przekracza 20% wariancji całkowitej (tab. 1.5), zatem zmienna ta zostanie wyłączona z analizy.

Benzo-a-piren

W przypadku oznaczeń tego wskaźnika w próbkach z obszaru RZGW Kraków w trakcie badań zmieniano wykonawcę badań i metodykę oznaczeń (zmiana granicy oznaczalności DL, Witczak et al., 1994), zatem zmienna ta będzie wyłączona z dalszej analizy. Ponieważ wszystkie uzyskane wyniki były poniżej granicy oznaczalności, nie będzie prowadzona też analiza statystyczna zbioru.

DDT

Średnie stężenie DDT w badanych próbkach wynosi $\bar{x} = 0.0000174$ mg/dm^3, przy zmienności w zakresie od 0.000003 do 0.000051 mg/dm^3 (współczynnik zmienności $V = 56.29\%$). Rozkład jest w przybliżeniu symetryczny (skośność: 0.207, kurtoza: -0.105).

Na podstawie wykresu typu „łodyga i liście" jedną obserwację zakwalifikowano jako odstającą (punkt 11043, wartość $x \geqslant 0.000051$ mg/dm^3). Po jej wyłączeniu z analizy współczynnik zmienności wynosi $V = 55.02\%$, wartość średnia $\bar{x} = 0.0000172$ mg/dm^3 a mediana jest równa $\tilde{x} = 0.00002$ mg/dm^3.

Ze względu na niską precyzję oznaczeń (75% obserwacji poniżej granicy oznaczalności DL, PDL/DL = 100 —- tab. 1.3), wysoki poziom wariancji technicznej (tab. 1.5), zmienna ta nie będzie uwzględniona w dalszej analizie.

DDE

Średnia zawartość DDE w badanym zbiorze wynosi $\bar{x} = 0.00000714$ mg/dm^3, przy zmienności w zakresie od 0.000005 do 0.000020 mg/dm^3 (współczynnik zmienności $V = 24.72\%$). Rozkład jest asymetryczny (skośność: 1.910) i silnie spiczasty (kurtoza: 15.503).

Z wykresu typu „łodyga i liście" wynika, że jedną obserwację należy zakwalifikować jako odstającą (punkt 21063; wartość $x \geqslant 0.00002$ mg/dm^3). Po jej wyłączeniu z analizy wartość średnia $\bar{x} = 0.00000706$ mg/dm^3, mediana wynosi $\tilde{x} = 0.000008$ mg/dm^3 a współczynnik zmienności $V = 20.61\%$.

Praktyczna granica oznaczalności PDL jest dwa rzędy wielkości wyższa od laboratoryjnej granicy oznaczalności DL. 92% wartości to wyniki poniżej granicy oznaczalności DL (tab. 1.3), wariancja techniczna σ^2_{tech} stanowi 25% wariancji całkowitej (tab. 1.5), zatem zmienna ta zostanie wyłączona z analizy.

DDD

Zbiór wartości stężeń DDD zawiera się w przedziale 0.000005–0.000100 mg/dm^3 (współczynnik zmienności $V = 116.31\%$), wartość średnia $\bar{x} = 0.000013$ mg/dm^3. Rozkład wartości charakteryzuje się asymetrią prawostronną (skośność: 2.991) i spiczastością (kurtoza: 9.287).

Z wykresu „łodyga i liście" wynika, że 18% obserwacji to obserwacje ekstremalne (trzydzieści obserwacji; wartości $x \geqslant 0.00002$ mg/dm^3). Tworzą one podzbiór obserwacji anomalnych o wartości średniej $\bar{x} = 0.00004$ mg/dm^3 i medianie $\tilde{x} = 0.00003$ mg/dm^3.

Po wyłączeniu z analizy obserwacji anomalnych współczynnik zmienności ma wartość $V = 26.88\%$. Średnie stężenie DDD w analizowanych próbkach kształtuje się na poziomie $\bar{x} = 0.0000076$ mg/dm^3, mediana $\tilde{x} = 0.000008$ mg/dm^3.

Wariancja techniczna stanowi wprawdzie mniej niż 20% zmienności całkowitej (tab. 1.5), jednak większość oznaczeń (55% obserwacji) to wyniki poniżej granicy oznaczalności DL. Ponadto praktyczna granica oznaczalności PDL jest trzy rzędy wielkości wyższa od DL (tab. 1.3), zatem zmienna ta zostanie wyłączona z analizy.

Gamma-HCH (lindan)

Średnia zawartość lindanu kształtuje się na poziomie $\bar{x} = 0.0000064$ mg/dm^3, przy zmienności w zakresie od 0.000003 do 0.000011 mg/dm^3 (współczynnik zmienności ma wartość $V = 36.47\%$). Rozkład jest lekko asymetryczny (skośność: -0.677) i spłaszczony (kurtoza: -1.323).

W zbiorze nie ma wartości odstających, wariancja techniczna stanowi 11% wariancji całkowitej (tab. 1.5), jednak przeważająca większość obserwacji (83% obserwacji) to wyniki poniżej granicy oznaczalności DL (tab. 1.3). Zmienna ta będzie wyłączona z analizy.

Metoksychlor

Zbiór wartości stężeń metoksychloru zawiera się w przedziale wartości od 0.000008 do 0.000603 mg/dm^3, przy średnim stężeniu na poziomie $\bar{x} = 0.0000391$ mg/dm^3. Współczynnik zmienności ma wartość $V = 123.7\%$. Rozkład wartości charakteryzuje się asymetrią prawostronną (skośność: 9.611) i spiczastością (kurtoza: 112.361).

Z wykresu typu „łodyga i liście" wynika, że jedną obserwację należy zakwalifikować jako ekstremalnie wysoką (punkt 11043, $x \geqslant 0.000603$ mg/dm^3). Po jej wyłączeniu z analizy współczynnik zmienności $V = 56.72\%$, wartość średnia $\bar{x} = 0.000036$ mg/dm^3, a mediana $\tilde{x} = 0.00005$ mg/dm^3.

Praktyczna granica oznaczalności PDL jest trzy rzędy wielkości wyższa od laboratoryjnej granicy oznaczalności DL (tab. 1.3). Wariancja techniczna σ^2_{tech} stanowi 30% wariancji całkowitej (tab. 1.5), 80% obserwacji to wyniki poniżej granicy oznaczalności DL, zatem zmienna ta nie będzie uwzględniona w dalszej analizie.

Suma 6WWA

Wskaźnik ten był oznaczany jedynie w próbkach pobranych z obszaru RZGW Katowice, zmienna ta nie będzie więc uwzględniana w dalszej analizie.

Średnie stężenie tego wskaźnika w dostępnym zbiorze 53 punktów (114 braków danych) kształtuje się na poziomie $\bar{x} = 0.000166$ mg/dm^3, przy rozrzucie wartości od 0.000022 do 0.000506 mg/dm^3. Współczynnik zmienności $V = 62.21\%$, rozkład jest w przybliżeniu symetryczny (skośność: 0.974, kurtoza: 0.941).

Ostatecznie w zweryfikowanej bazie danych, na podstawie której będą prowadzone próby prognozowania jakości wód podziemnych w układzie przestrzennym pozostało szesnaście zmiennych (tab. 1.13):

- temperatura [°C];
- odczyn pH;
- suma substancji rozpuszczonych [mg/dm^3];
- zasadowość ogólna [mval/dm^3];
- twardość ogólna [mg $CaCO_3$/dm^3];
- sód [mg/dm^3];
- magnez [mg/dm^3];
- wapń [mg/dm^3];
- chlorki [mg/dm^3];
- siarczany [mg/dm^3];
- krzemionka zdysocjowana [mg/dm^3];
- fluorki [mg/dm^3];
- cynk [mg/dm^3];
- współczynnik absorpcji UV (A 254);
- rozpuszczony węgiel organiczny [mg/dm^3];
- utlenialność ChZT-Mn [mg/dm^3].

W tabeli 1.13 zestawiono charakterystykę zmiennych w zweryfikowanym zbiorze danych wraz z wartością średnią $\bar{x}$ oraz parametry, na podstawie których dokonywano weryfikacji zmiennych.

Tabela 1.13. Charakterystyka zmiennych w zweryfikowanym zbiorze danych dla zlewni górnej Wisły. Objaśnienia: N — liczba obserwacji; B — liczba braków danych; $\bar{x}$ — wartość średnia [mg/dm^3]; DL — laboratoryjna granica oznaczalności; PDL — praktyczna granica oznaczalności; σ^2_{tech} — wariancja techniczna jako procent wariancji całkowitej, oszacowana metodą klasycznej analizy wariancji ANOVA; znak — oznacza brak danych

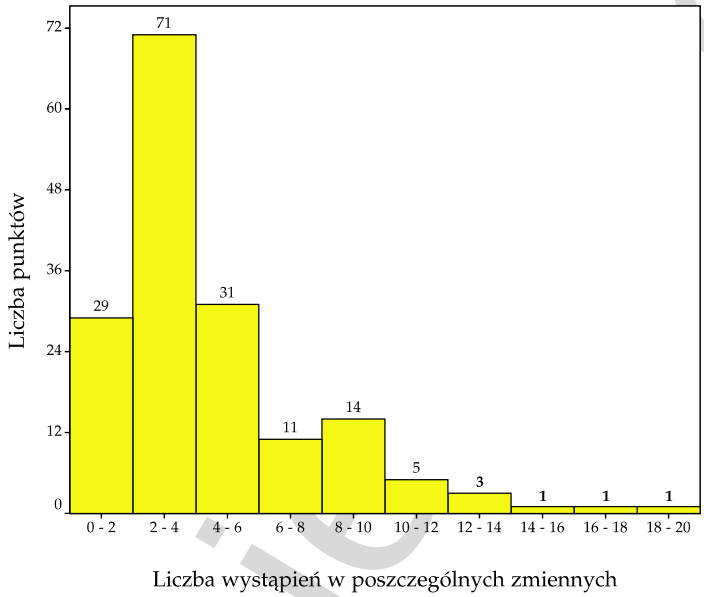
Lp.	Analizowana zmienna	Jednostka	N	B	$\bar{x}$	PDL/DL	$\bar{x}$/DL	σ^2_{tech}
1.	Temperatura	°C	160	7	9.98	—	—	9.39
2.	Odczyn pH		156	11	7.15	—	—	8.33
3.	Suma substancji rozpuszczonych	mg/dm^3	163	4	346.82	4	350	5.36
4.	Zasadowość ogólna	mval/dm^3	166	1	4.03	3	80	3.47
5.	Twardość ogólna	mg $CaCO_3$/dm^3	165	2	268.32	1	135	3.61
6.	Sód	mg/dm^3	155	12	7.70	1	77	15.39
7.	Magnez	mg/dm^3	162	5	14.18	7	142	18.04
8.	Wapń	mg/dm^3	165	2	76.09	48	761	4.09
9.	Chlorki	mg/dm^3	157	10	20.40	1	4	15.36
10.	Siarczany	mg/dm^3	160	7	43.85	1	4	2.04
11.	Krzemionka zdysocjowana	mg/dm^3	166	1	12.44	1	18	4.45
12.	Fluorki	mg/dm^3	158	9	0.19	1	2	3.54
13.	Cynk	mg/dm^3	146	21	0.05	7.5	5	86.85
14.	Współczynnik absorpcji UV (A 254)		149	18	0.08	6.4	17	8.37
15.	Rozpuszczony węgiel organiczny	mg/dm^3	153	14	1.33	12	7	12.86
16.	Utlenialność ChZT-Mn	mg/dm^3	153	14	1.45	2	3	5.35

Stosunek PDL/DL pozwala na porównanie praktycznej granicy oznaczalności z granicą oznaczalności podawaną przez laboratoria. Praktyczna granica oznaczalności PDL powinna być jak najbliższa laboratoryjnej granicy oznaczalności DL. W idealnym przypadku PDL = DL

Wartość $\bar{x}/\mathrm{DL}$ umożliwia odniesienie uzyskiwanych wyników pomiarów do laboratoryjnej granicy oznaczalności. Wyniki oznaczeń cechują się zadowalającą precyzją, jeśli są 1–2 rzędy wielkości wyższe od granicy oznaczalności DL, czyli jeśli $\bar{x}/\mathrm{DL} \geqslant 10$.

Wariancja techniczna σ^2_{tech} jest czynnikiem umożliwiającym ocenę precyzji oznaczeń wskaźników fizyko-chemicznych wód, nie może przekraczać 20% wariancji całkowitej σ^2_{tot}.

Podsumowano częstość, z jaką punkty monitoringowe były klasyfikowane jako obserwacje anomalne (we wszystkich analizowanych zmiennych — rys. 1.28).



Rysunek 1.28. Częstość, z jaką punkty monitoringowe były klasyfikowane jako obserwacje anomalne

Największą liczbę razy w grupie tej znalazły się punkty leżące w północnej części obszaru RZGW Katowice: punkt 11014 (dwadzieścia razy) i punkt 11013 (siedemnaście razy).

Są to punkty znajdujące się w chodnikach wodnych nieczynnej kopalni rud Zn–Pb „Orzeł biały", zbierających wodę z dużego obszaru (Witczak et al., 1994; Witkowski, 1997).

1.3. Analiza geostatystyczna i wyznaczenie naturalnego tła hydrogeochemicznego zweryfikowanych wskaźników fizyko-chemicznych

Dla danych zweryfikowanych w oparciu o parametry kontroli jakości (tab. 1.13) wykonano ocenę geostatystyczną (metodą krigingu) za pomocą programu GEO-EAS v. 1.2.1 (Englund, Sparks, 1991).

W metodzie tej zmienność pola hydrogeochemicznego oraz błędy ceny parametrów mogą zostać oszacowane na podstawie funkcji ujmującej zależność między średnim zróżnicowaniem parametrów a odległością między punktami ich pomiarów.

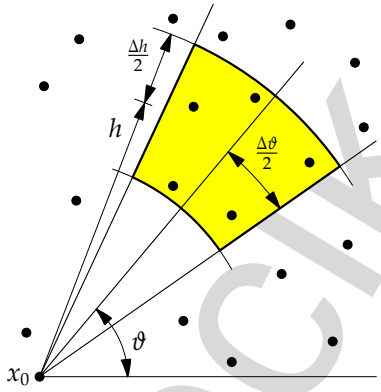
Funkcja ta, zwana semiwariogramem, pozwala określić model zmienności pola hydrogeochemicznego i oszacować zgodny z tym modelem przestrzenny rozkład badanego parametru (Macioszczyk, Witczak, 1999; Kania 2000).

Dla dyskretnej i regularnej siatki pomiarów postać semiwariogramu określana jest na podstawie klasycznego wzoru Matherona (Mucha, 1994):

$$\gamma(h) = \frac{1}{2n_h} \sum_{i=1}^{n_h} \left(z_{\vec{h}+1} - z_i \right)^2 \tag{1.9}$$

gdzie: z_i, $z_{\vec{h}+1}$ — wartości parametru w punktach oddalonych o wektor $\vec{h}$; n_h — liczba par punktów pomiarowych odległych o wektor $\vec{h}$.

W przypadku regionalnego monitoringu wód podziemnych sieć pomiarów ma charakter nieregularny, wówczas technika obliczeniowa jest bardziej skomplikowana, ma charakter przybliżony. Punkty pomiarowe grupowane są wtedy w klasy odległości Δh i przedziały kątowe zliczania $\Delta\vartheta$ względem każdego z punktów pomiarowych osobno (rys. 1.29).



Rysunek 1.29. Sposób obliczania semiwariogramu empirycznego dla dwuwymiarowej nieregularnej sieci opróbowania. Objaśnienia: x_0 — punkt bazowy z przypisaną wartością parametru; • — punkty opróbowań z przypisanymi wartościami parametru; żółty zacieniowany obszar oznacza sektor zliczania danych (Mucha, 1994)

Następnie obliczany jest średni kwadrat różnic dla wszystkich par utworzonych z wartości parametru określonych w punkcie wyróżnionym (bazowym) x_0 i w każdym z punktów, który znalazł się w obszarze grupowania danych.

Semiwariogram ustala się przypisując średniemu kwadratowi różnic średnią odległość między punktem x_0 a punktami z rozpatrywanego sektora zliczania. Następnie procedura jest powtarzana dla kolejnego przedziału odległości i dalej, dla kolejnych punktów pomiarowych, które przejmują rolę punktów bazowych (Mucha, 1994). W dalszej analizie przybliża się semiwariogram empiryczny jedną z funkcji analitycznych, które w dalszym postępowaniu traktowane są jako geostatystyczne modele zmienności.

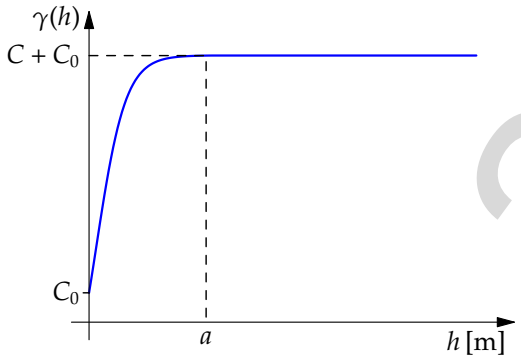
Mucha (1994) wyróżnia trzy typy modeli teoretycznych:

- model bez asymptoty (liniowy, de Wijsa, potęgowy) — charakteryzujący się nieograniczonym teoretycznie wzrostem semiwariogramu;
- model z asymptotą (Matherona, sferyczny — rys. 1.30, liniowy, Formery'ego, Gaussa) — z ograniczonym wzrostem semiwariogramu;
- model losowy — opisujący czysto losowe zróżnicowanie wartości parametru (brak autokorelacji między obserwacjami).

Aproksymacja semiwariogramów empirycznych może odbywać się metodą automatycznego dopasowania (metoda najmniejszych kwadratów) lub metodą dopasowania graficznego (ocena wizualna).

W celu stwierdzenia lub wykluczenia anizotropii zmienności badanego wskaźnika porównuje się przebiegi semiwariogramów kierunkowych przy różnych orientacjach pasów zliczeniowych. Jeśli zróżnicowanie to jest niewielkie, przyjmuje się, że struktura zmienności

parametru jest izotropowa i do dalszych obliczeń wykorzystuje się semiwariogram uśredniony (Mucha, 1994). Następnie wykorzystuje się procedurę krigingu do oszacowania średnich wartości analizowanego parametru i ich wartości w punktach analizowanego obszaru na podstawie znajomości struktury zmienności wyrażonej semiwariogramem.



Rysunek 1.30. Semiwariogram dla modelu sferycznego (Mucha 1994). Objaśnienia: $C + C_0$ — amplituda semiwariogramu (wartość, którą semiwariogram osiąga dla $h = a$); C_0 — stała efektu samorodków (wartość charakteryzująca zmienność lokalną badanego parametru i odpowiadająca składnikowi losowemu zmienności dla $h \to 0$); a — zasięg semiwariogramu (odległość h, powyżej której semiwariogram nie zwiększa już swej wartości).

W porównaniu z innymi procedurami interpolacyjnymi metoda krigingu charakteryzuje się większą dokładnością ze względu na minimalizację błędu oceny parametru.

Dla poszczególnych punktów siatki interpolacyjnej lub bloków obliczeniowych wyznaczone zostają:

- wartość średnia parametru:

$$z_k^* = \sum_{i=1}^{n} w_{ik} \cdot z_i \tag{1.10}$$

  gdzie: w_{ik} — współczynnik wagowy krigingu, zależny od wielkości bloku, dla którego dokonuje się oszacowania średniej wartości parametru, od odległości między punktami pomiaru oraz od autokorelacji między obserwacjami określonej przez model wariogramu; z_i — wartość parametru w i-tym punkcie pomiarowym;
- standardowy błąd bezwzględny oszacowania wartości średniej parametru σ_k, którego wariancja opisana jest wzorem:

$$\sigma_k^2 = \sum_{i=1}^{n} w_{ik} \cdot \overline{\gamma}(S_i, A) + \lambda - \overline{\gamma}(A, A) \tag{1.11}$$

  gdzie: $\overline{\gamma}(S_i, A)$ — średnia wartość semiwariogramu dla wszystkich możliwych odcinków łączących punkty pomiaru z ocenianym obszarem; λ — mnożnik Lagrange'a; $\overline{\gamma}(A, A)$ — średnia wartość semiwariogramu dla wszystkich odcinków, których końce zawarte są w obrębie ocenianego obszaru;
- standardowy błąd względny oszacowania wartości średniej parametru σ_{kw}:

$$\sigma_{kw} = \frac{\sigma_k}{z_k^*} \tag{1.12}$$

W efekcie uzyskuje się mapy izoliniowe rozkładu poszczególnych wskaźników oraz mapy błędów interpolacji świadczące o wiarygodności mapy rozkładu.

Na rysunkach 1.31–1.46 przedstawiono rezultaty obliczeń geostatystycznych dla analizowanych wskaźników (mapy izoliniowe rozkładu analizowanych wskaźników oraz mapy błędów oszacowania wraz z parametrami modelu).

Błąd względny oszacowania rozkładu wskaźników fizyko-chemicznych z wykorzystaniem modelowania geostatystycznego w większości przypadków nie przekracza 40%. Obserwuje się zależność tego błędu od gęstości opróbowania danego obszaru — mniejszy błąd względny oszacowania w przypadku gęsto położonych punktów RMWP (np. w obszarze RZGW Katowice).

Z map tych wynika również, że w zlewni górnej Wisły nie ma obszarów różniących się pod względem stężeń analizowanych wskaźników (brak regionów anomalnych), zatem można wyznaczyć regionalne tło hydrogeochemiczne dla całego badanego obszaru.

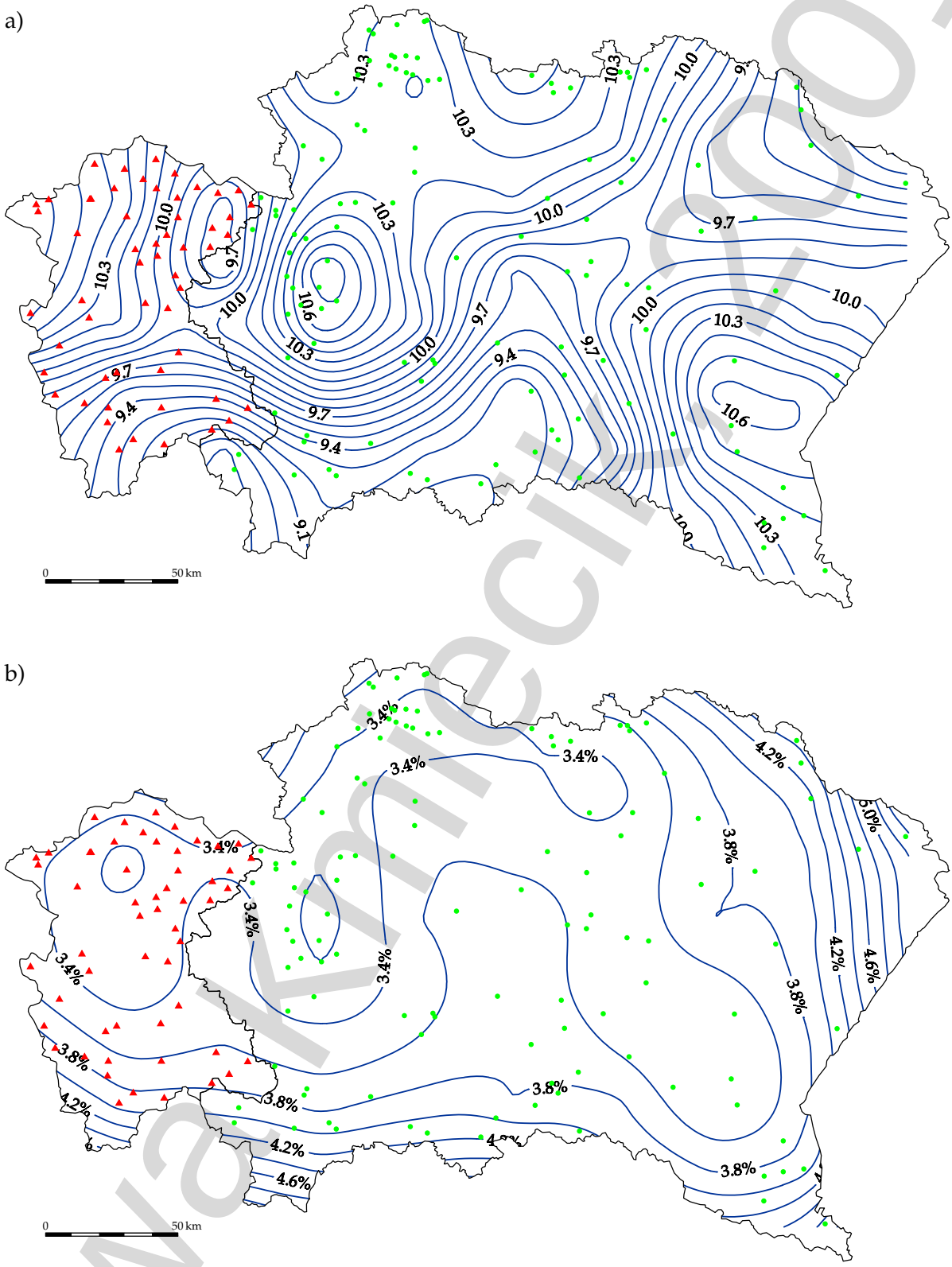
Tabela 1.14. Porównanie tła hydrogeochemicznego wód podziemnych w zlewni górnej Wisły (RZGW Katowice i RZGW Kraków) z ogólnopolskimi danymi (Witczak, Adamczyk, 1994) i standardami dla wód pitnych (Dz.U. Nr 82 z 2000 r., poz. 937, Dyrektywa Unii Europejskiej 98/83/EC, wytyczne WHO, 1998). Objaśnienia: MDP — maksymalne dopuszczalne stężenie analizowanego wskaźnika w wodach pitnych; znak — oznacza brak danych

Lp.	Wskaźnik	Jednostka	Tło hydrogeochemiczne		MDP
			zlewnia górnej Wisły	ogólnopolskie	
1.	Temperatura	°C	9–11	4–20	—
2.	Odczyn pH		6.6–7.6	6.5–8.5	6.5–9.5
3.	Suma substancji rozpuszczonych	mg/dm^3	162–485	200–500	—
4.	Zasadowość ogólna	$mval/dm^3$	2–6	1–6	—
5.	Twardość ogólna	mg $CaCO_3/dm^3$	133–400	100–400	—
6.	Sód	mg/dm^3	1.9–14.3	1–60	200
7.	Magnez	mg/dm^3	3.8–27.3	0.5–50	50
8.	Wapń	mg/dm^3	37.4–113.2	2–200	—
9.	Chlorki	mg/dm^3	5.7–39.7	2–60	250
10.	Siarczany	mg/dm^3	15–84	5–60	250
11.	Krzemionka zdysocjowana	mg/dm^3	4–22	1–30	—
12.	Fluorki	mg/dm^3	0.03–0.33	0.05–0.50	1.5
13.	Cynk	mg/dm^3	0.01–0.09	0.05–0.50	3
14.	Współczynnik absorpcji UV (A 254)		0.02–0.16	0.01–0.50	—
15.	Rozpuszczony węgiel organiczny	mg/dm^3	0.7–2.2	1–10	—
16.	Utlenialność ChZT-Mn	mg/dm^3	0.8–2.3	1–10	5000

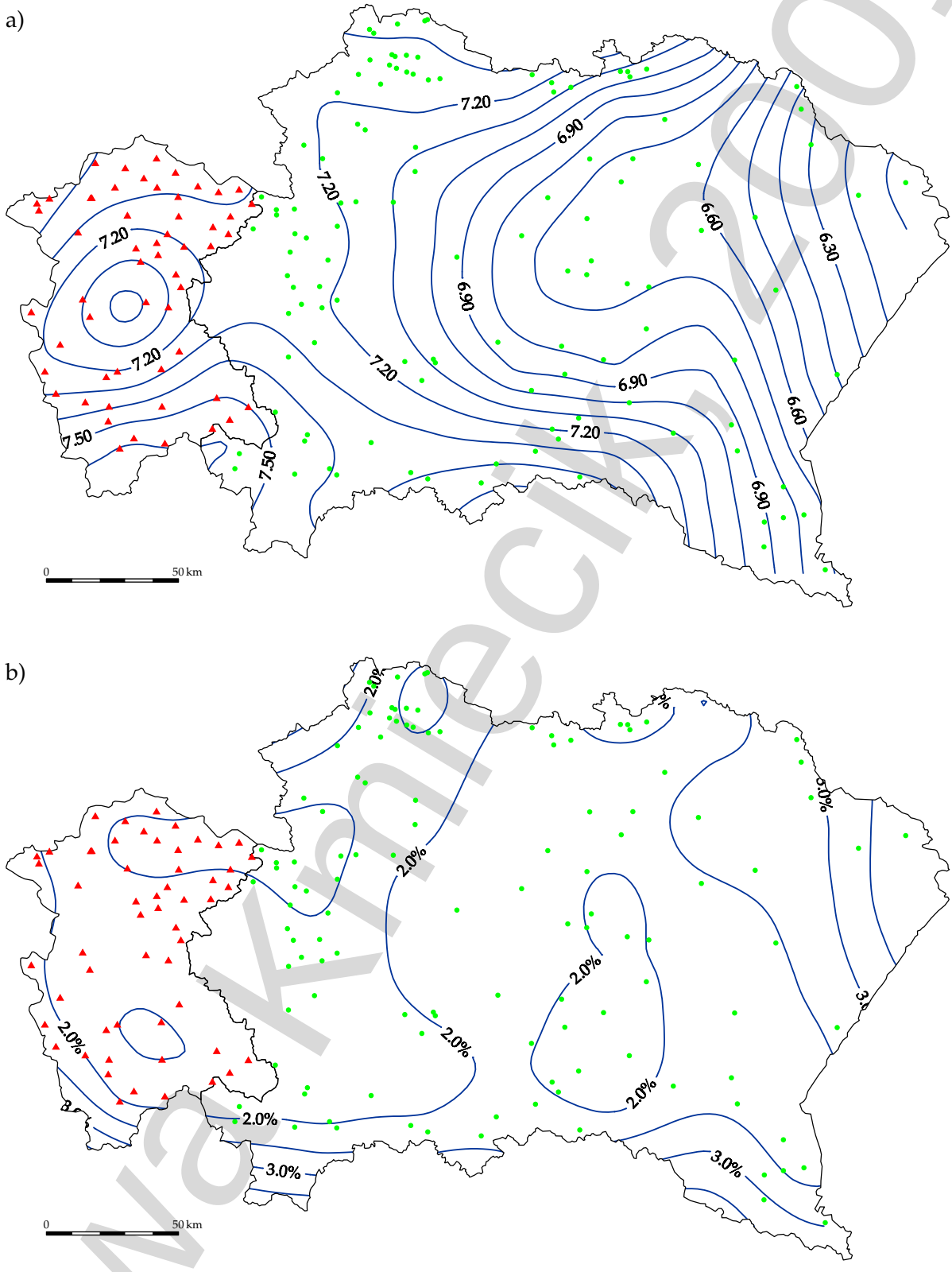
Tło regionalne w zlewni górnej Wisły wyznaczono zgodnie z metodą zaproponowaną przez Kanię (2000). Zgodnie z tą metodą zakres tła hydrogeochemicznego wyznacza się ze wzoru $\bar{z} \pm 1\mathrm{Sd}$ (gdzie $\bar{z}$ to mediana, a Sd — odchylenie standardowe), co oznacza, że dolną granicę tła stanowi 16, a górną 84 percentyl danych (Kania, 2000).

W tabeli 1.14 zestawiono uzyskane wartości regionalnego tła hydrogeochemicznego dla obszaru dorzecza górnej Wisły wraz z wartościami tła ogólnopolskiego (Macioszczyk, 1990; Witczak, Adamczyk, 1994) i maksymalnymi dopuszczalnymi stężeniami analizowanych wskaźników w wodach pitnych (Dz.U. Nr 82 z 2000 r., poz. 937; Dyrektywa Unii Europejskiej 98/83/EC; wytyczne WHO, 1998).

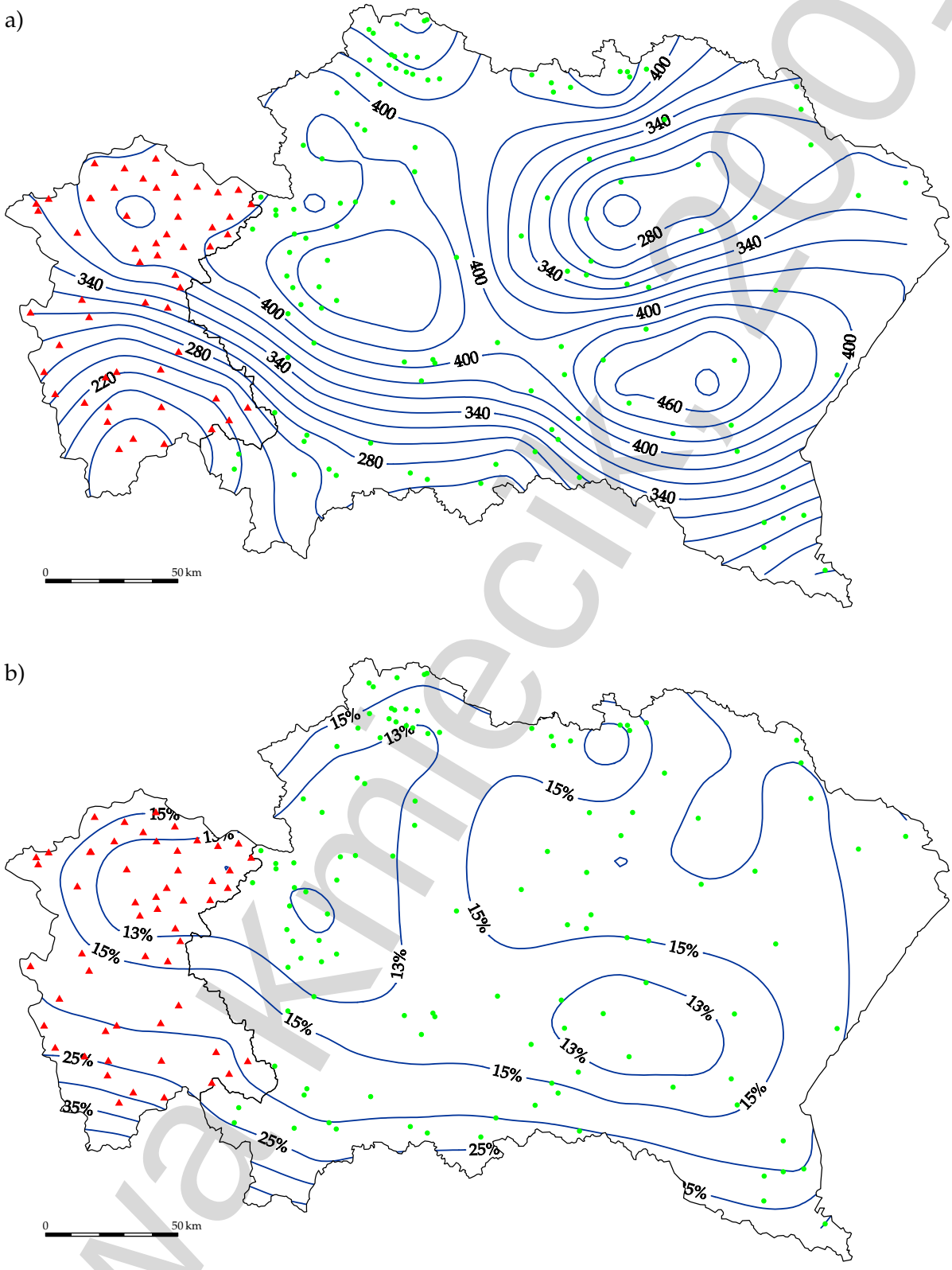
Porównując uzyskane wartości regionalnego tła hydrogeochemicznego obszaru zlewni górnej Wisły z wartościami ogólnopolskimi, należy stwierdzić, że jedynie w przypadku siarczanów uzyskano wartości wyższe od ogólnopolskiego tła hydrogeochemicznego (Witczak, Adamczyk, 1994), co może świadczyć o niekorzystnym wpływie aglomeracji przemysłowej na wody podziemne omawianego obszaru.



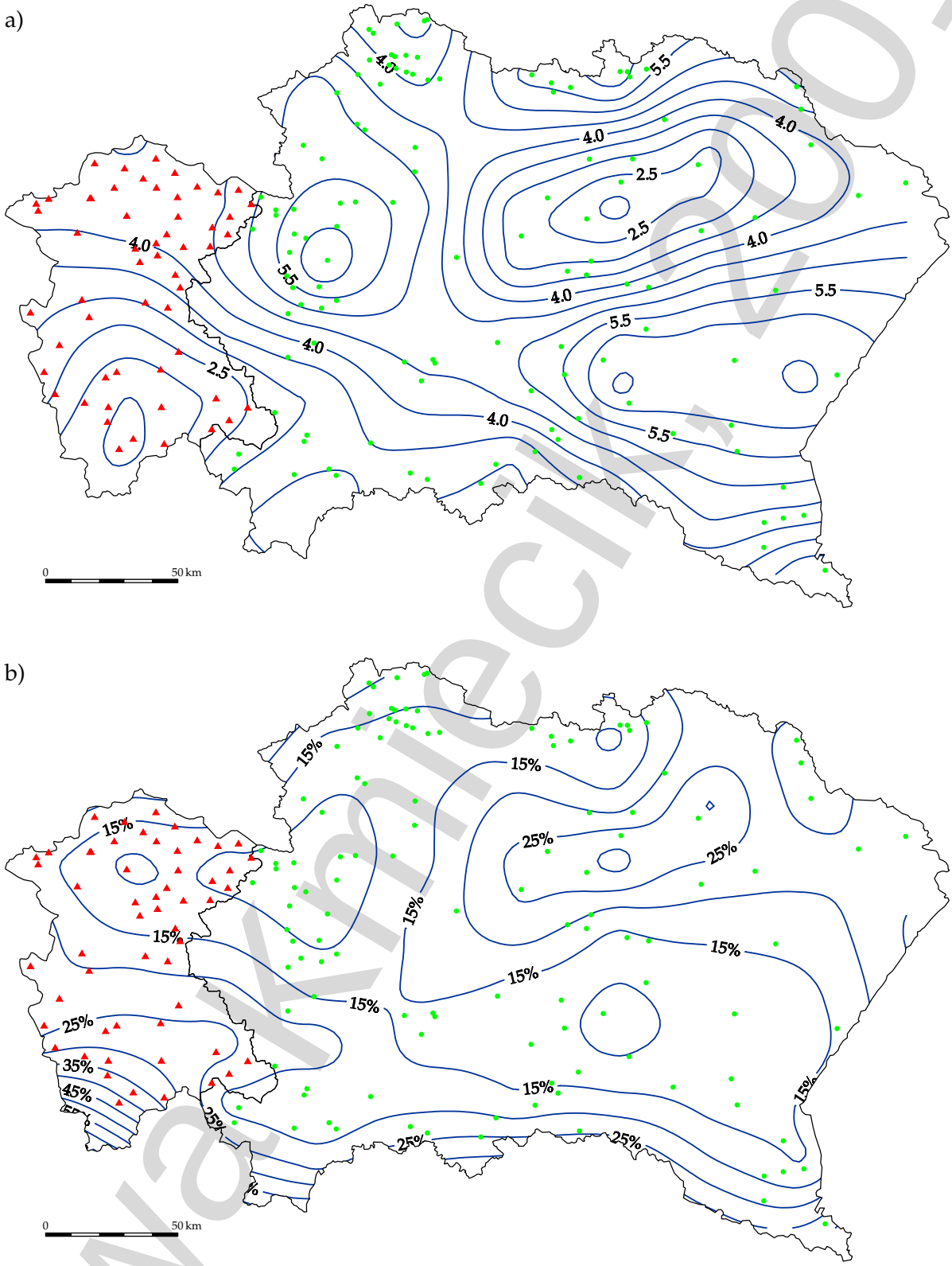
Rysunek 1.31. Mapa izoliniowa rozkładu temperatury [°C] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą kringu blokowego wg modelu liniowego o parametrach $C_0 = 1.6$, $C = 0.1$, $a = 50\,000$ m. Opróbowanie: 1993 rok, seria 1



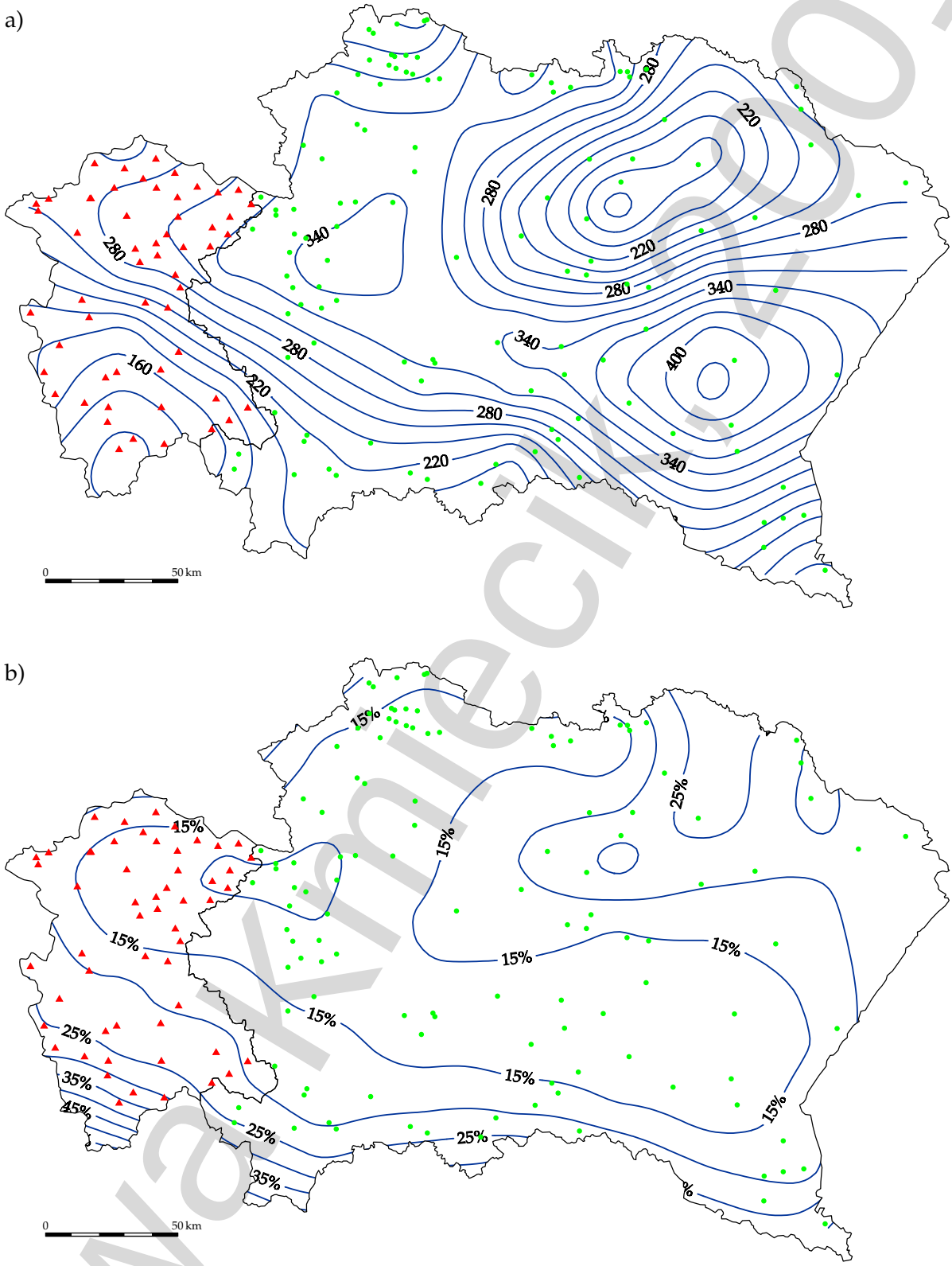
Rysunek 1.32. Mapa izoliniowa rozkładu pH w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu liniowego o parametrach $C_0 = 0.075$, $C = 0.09$, $a = 50\,000$ m. Opróbowanie: 1993 rok, seria 1



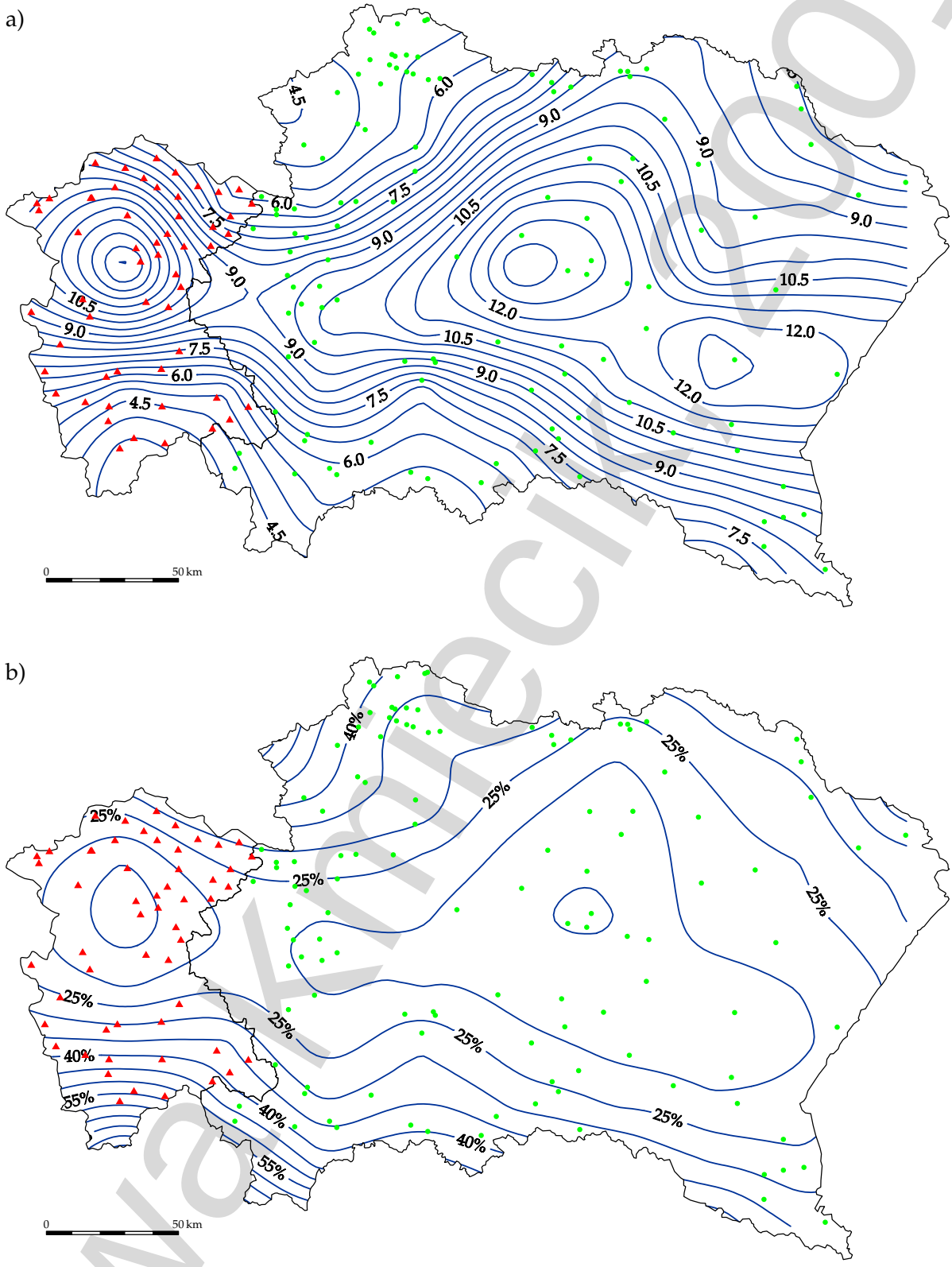
Rysunek 1.33. Mapa izoliniowa rozkładu sumy substancji rozpuszczonych [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu sferycznego o parametrach $C_0 = 17000$, $C = 10000$, $a = 85\,000$ m. Opróbowanie: 1993 rok, seria 1



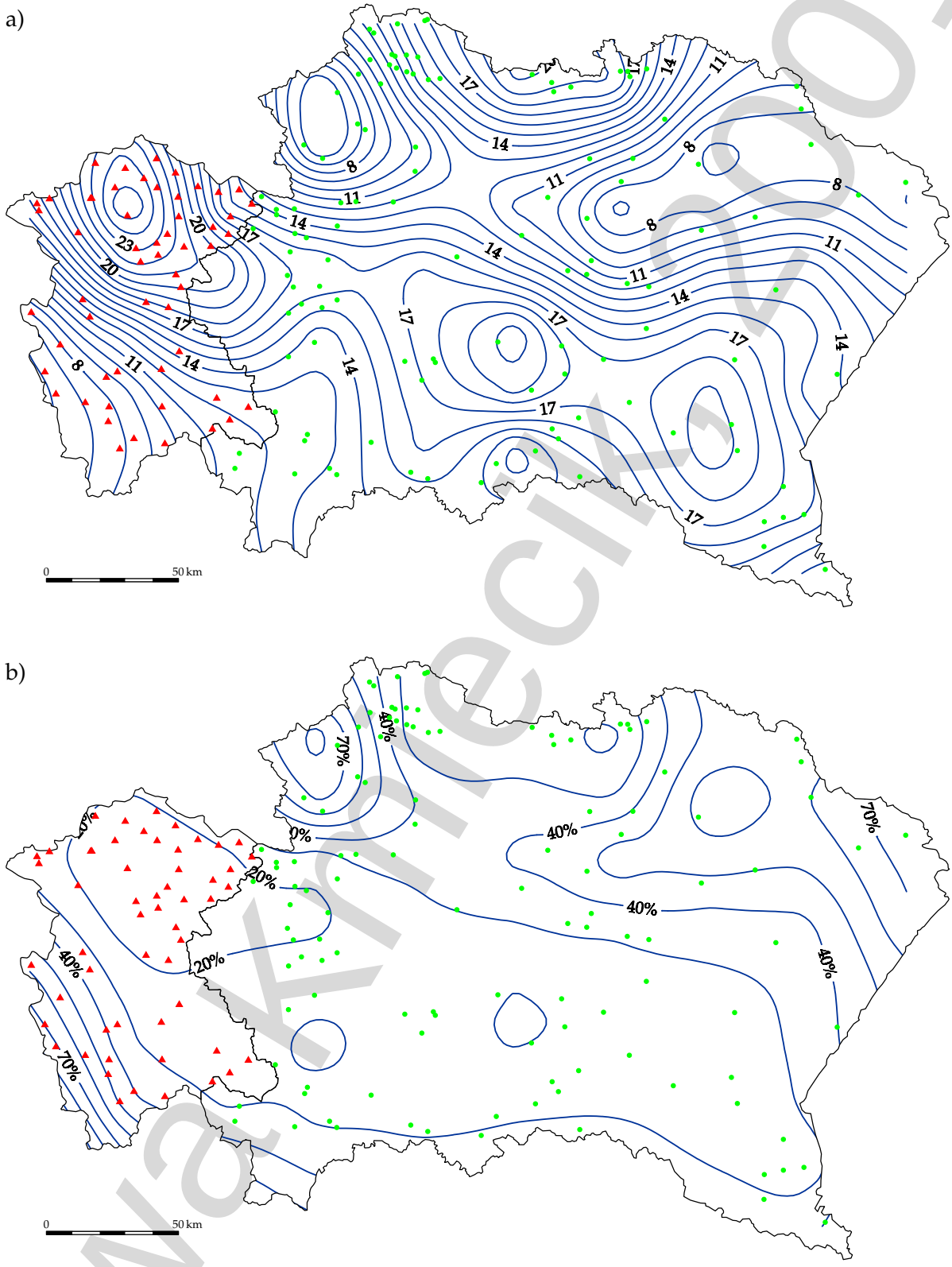
Rysunek 1.34. Mapa izoliniowa rozkładu zasadowości ogólnej [mval/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu sferycznego o parametrach $C_0 = 0.8$, $C = 3.1$, $a = 80\,000$ m. Opróbowanie: 1993 rok, seria 1



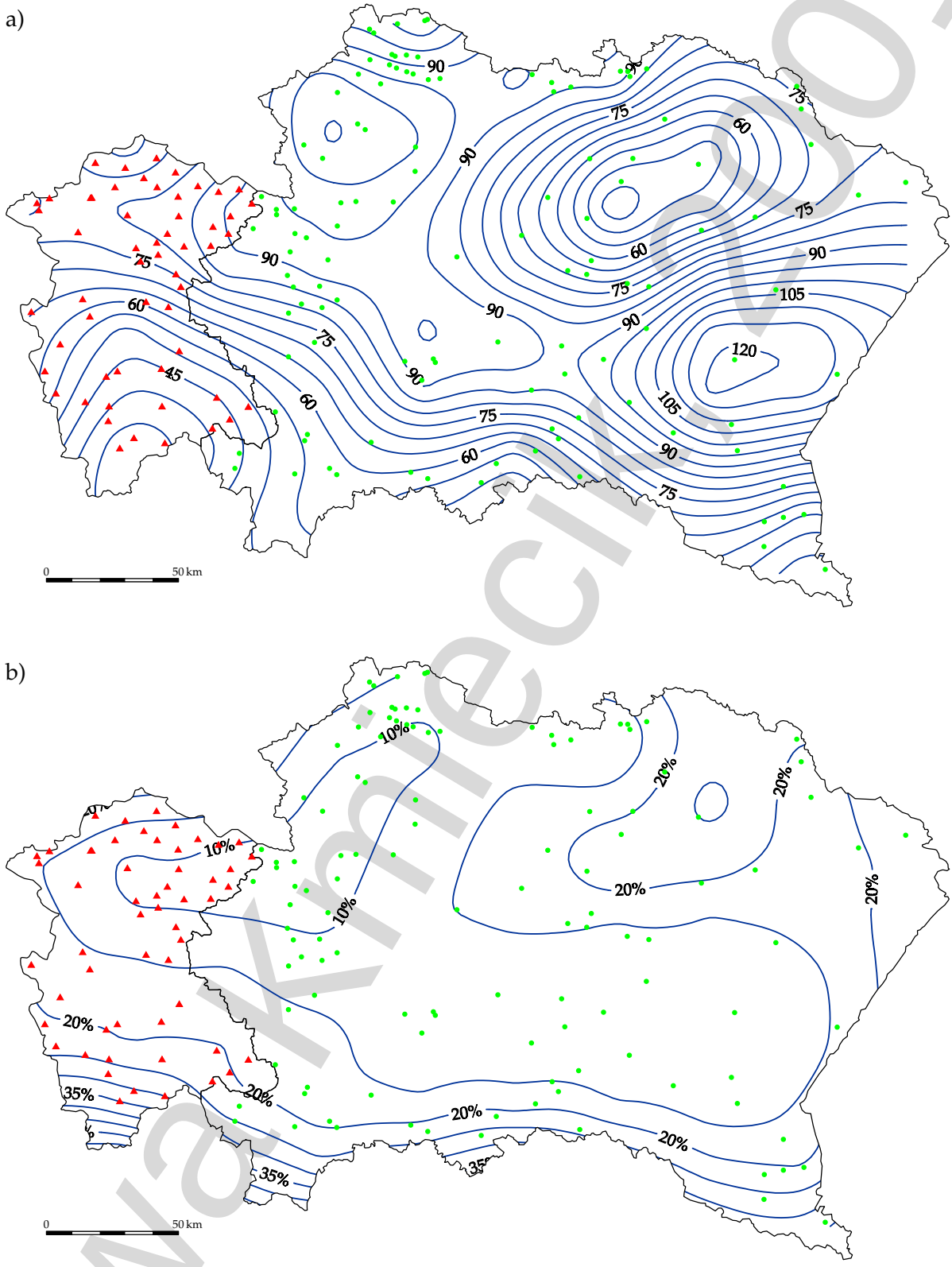
Rysunek 1.35. Mapa izoliniowa rozkładu twardości ogólnej [mg $CaCO_3/dm^3$] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu sferycznego o parametrach $C_0 = 7000$, $C = 8500$, $a = 75\,000$ m. Opróbowanie: 1993 rok, seria 1



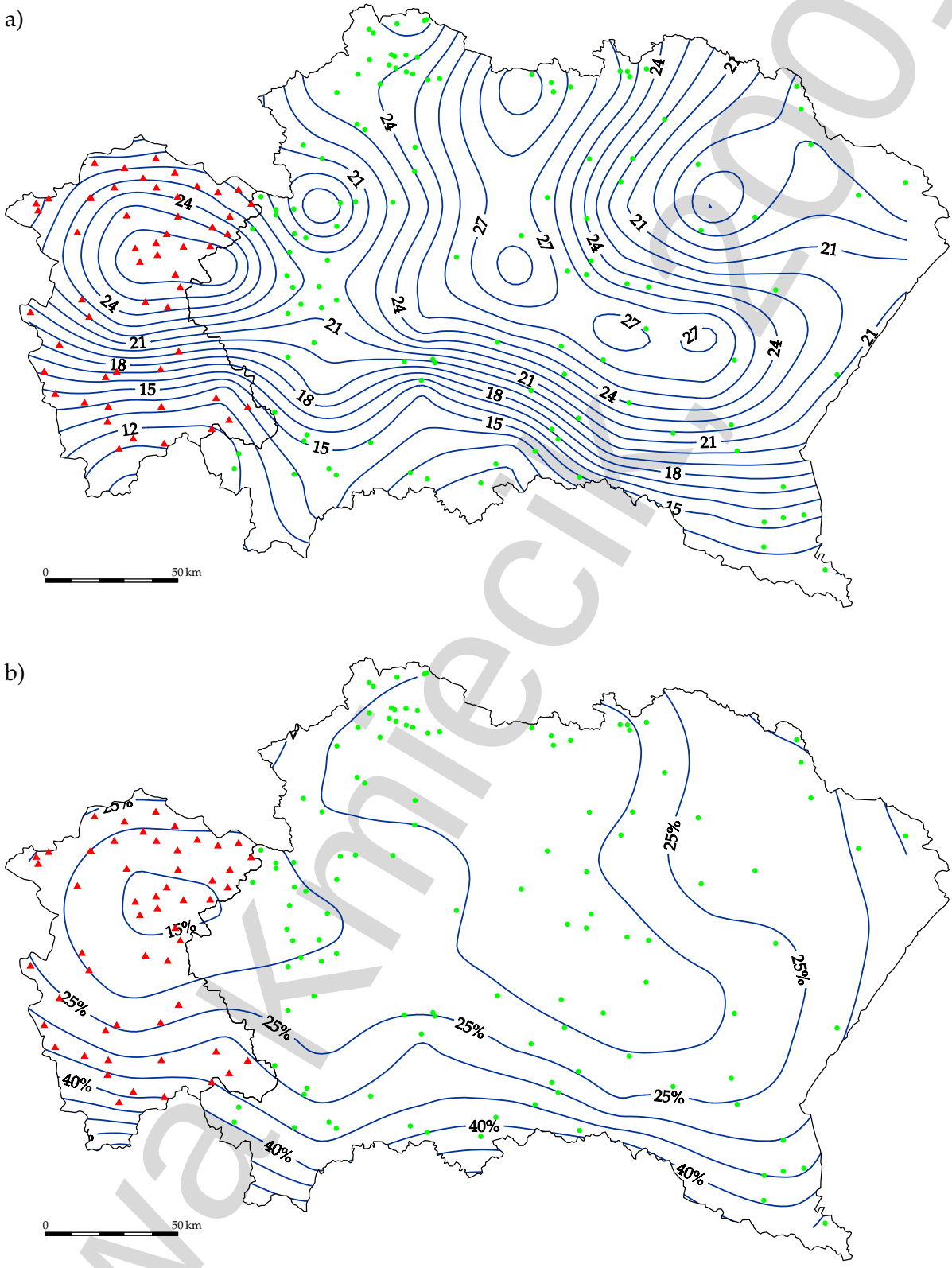
Rysunek 1.36. Mapa izoliniowa rozkładu sodu [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą kringingu blokowego wg modelu sferycznego o parametrach $C_0 = 35$, $C = 8$, $a = 80\,000$ m. Opróbowanie: 1993 rok, seria 1



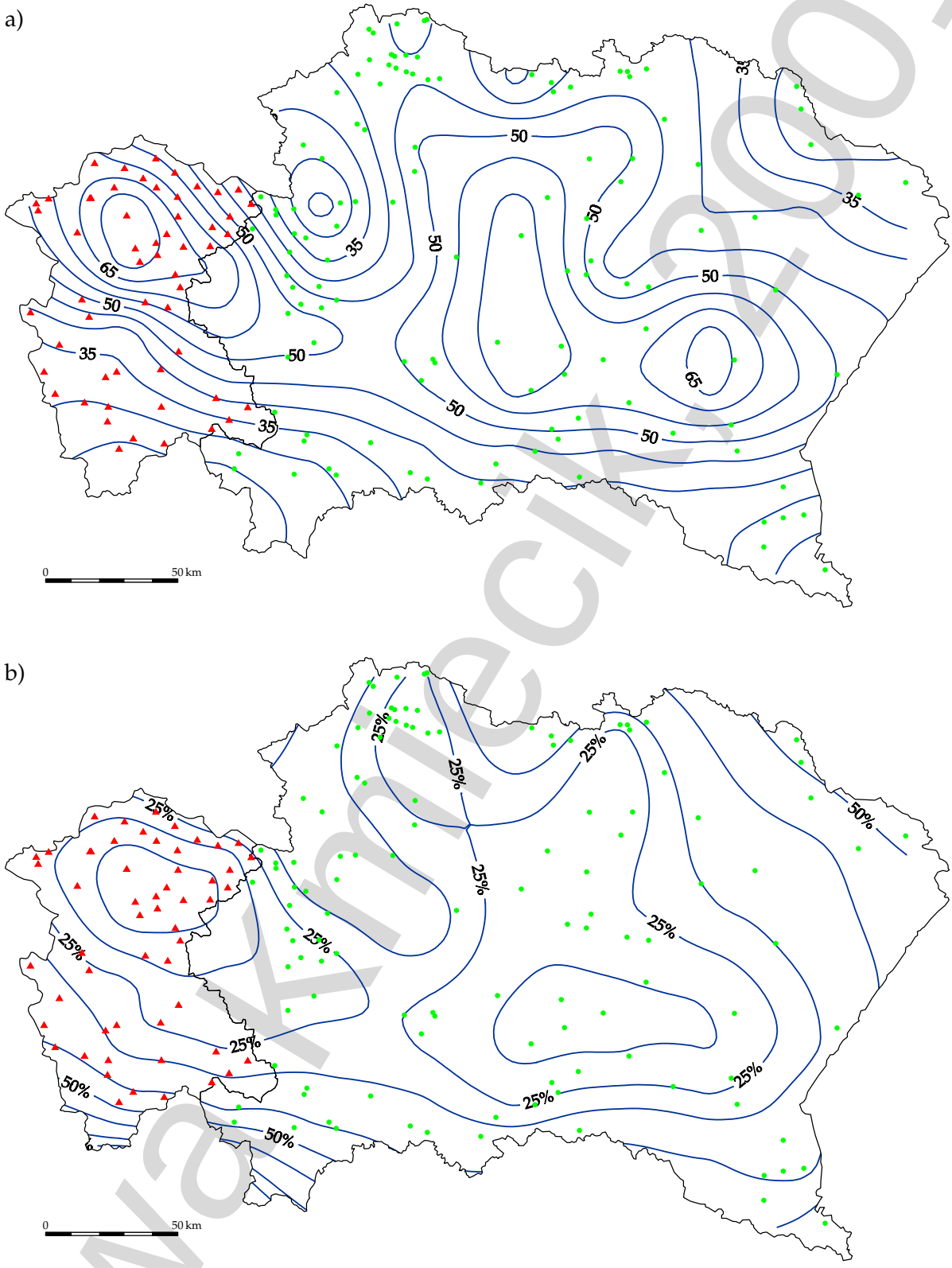
Rysunek 1.37. Mapa izoliniowa rozkładu magnezu [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą kriging blokowego wg modelu sferycznego o parametrach $C_0 = 60$, $C = 55$, $a = 60\,000$ m. Opróbowanie: 1993 rok, seria 1



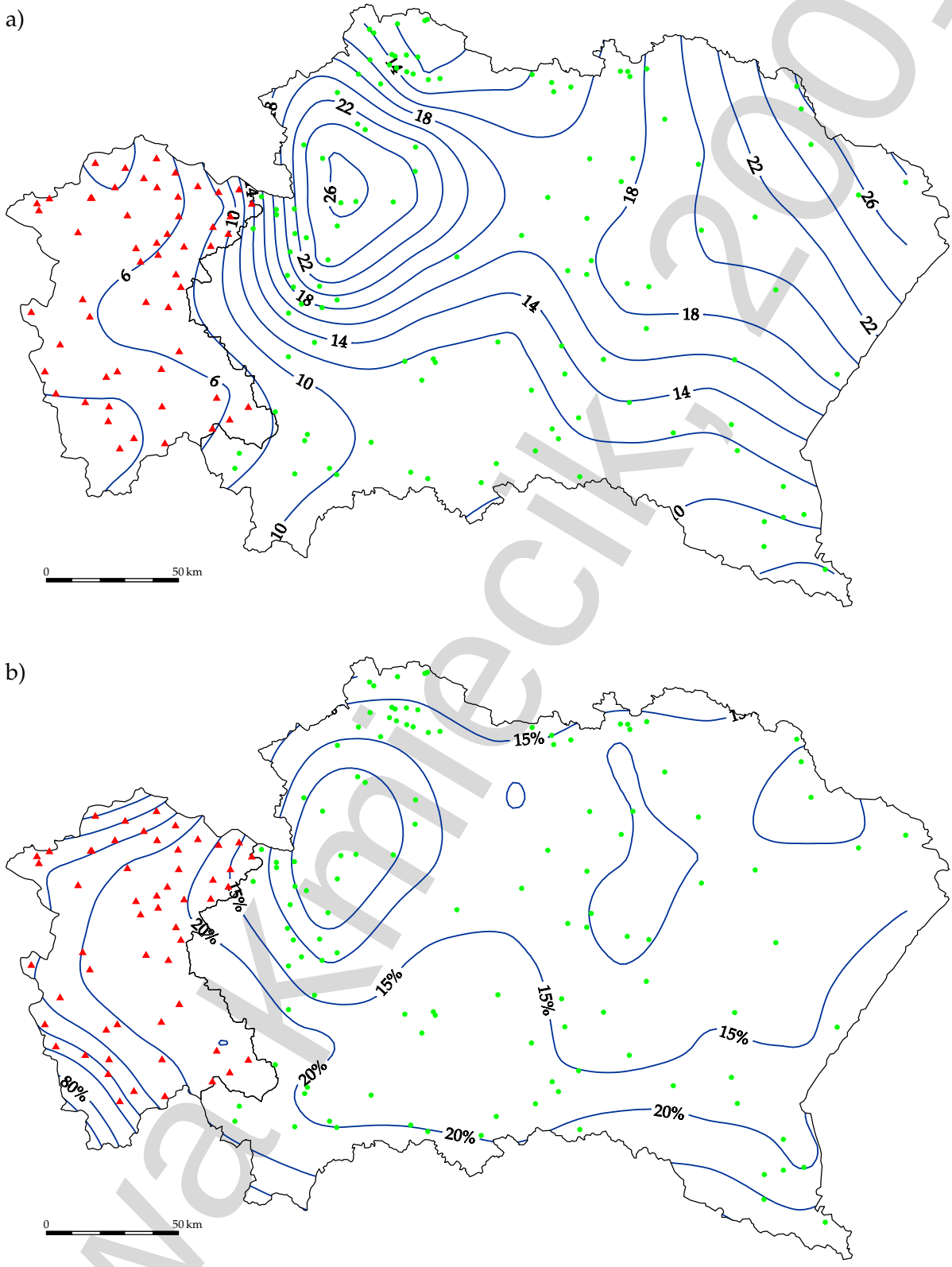
Rysunek 1.38. Mapa izoliniowa rozkładu wapnia [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą kriggingu blokowego wg modelu Gaussa o parametrach $C_0 = 700$, $C = 750$, $a = 80\,000$ m. Opróbowanie: 1993 rok, seria 1



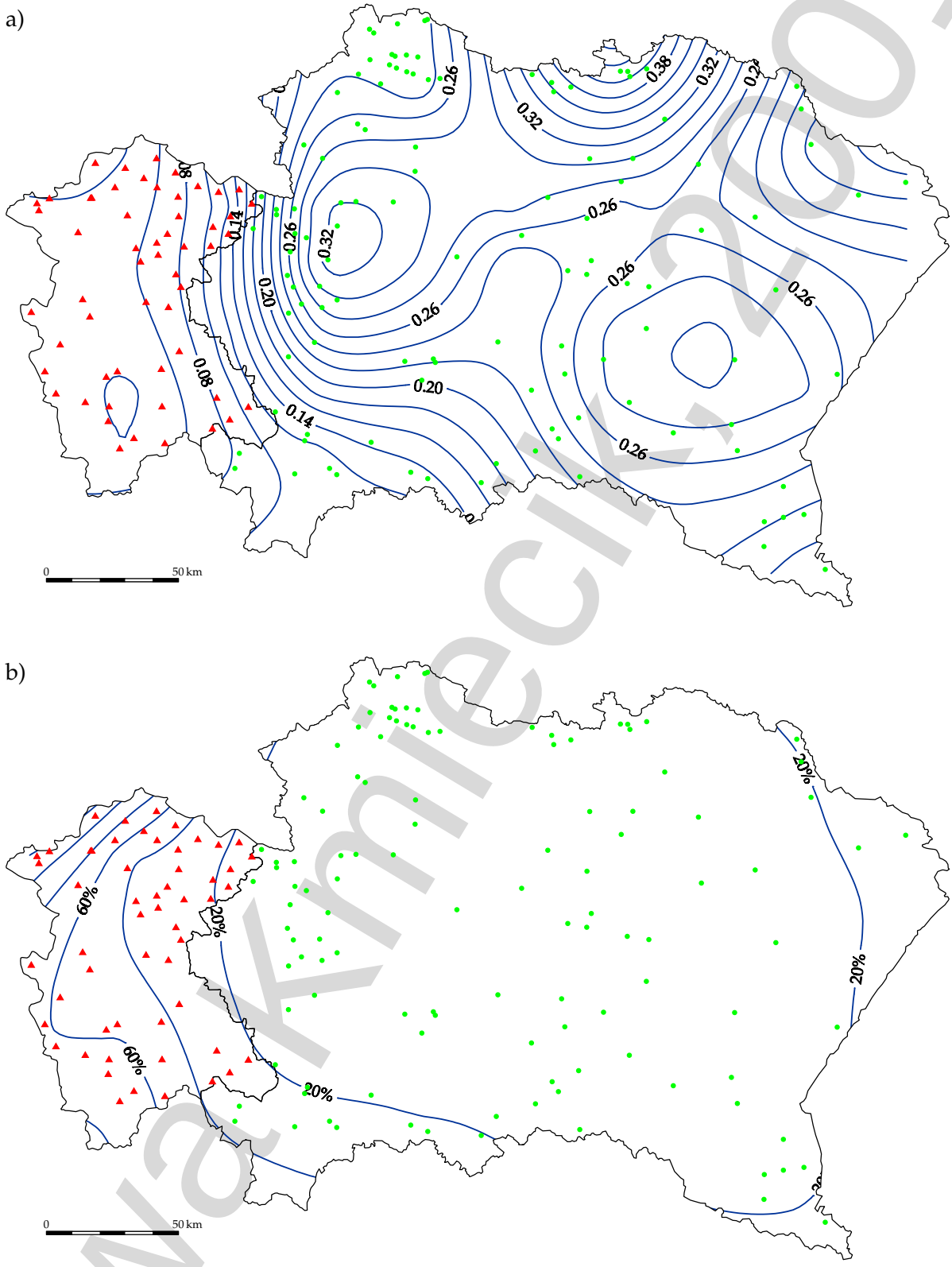
Rysunek 1.39. Mapa izoliniowa rozkładu chlorków [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą kriging blokowego wg modelu Gaussa o parametrach $C_0 = 205$, $C = 35$, $a = 60\,000$ m. Opróbowanie: 1993 rok, seria 1



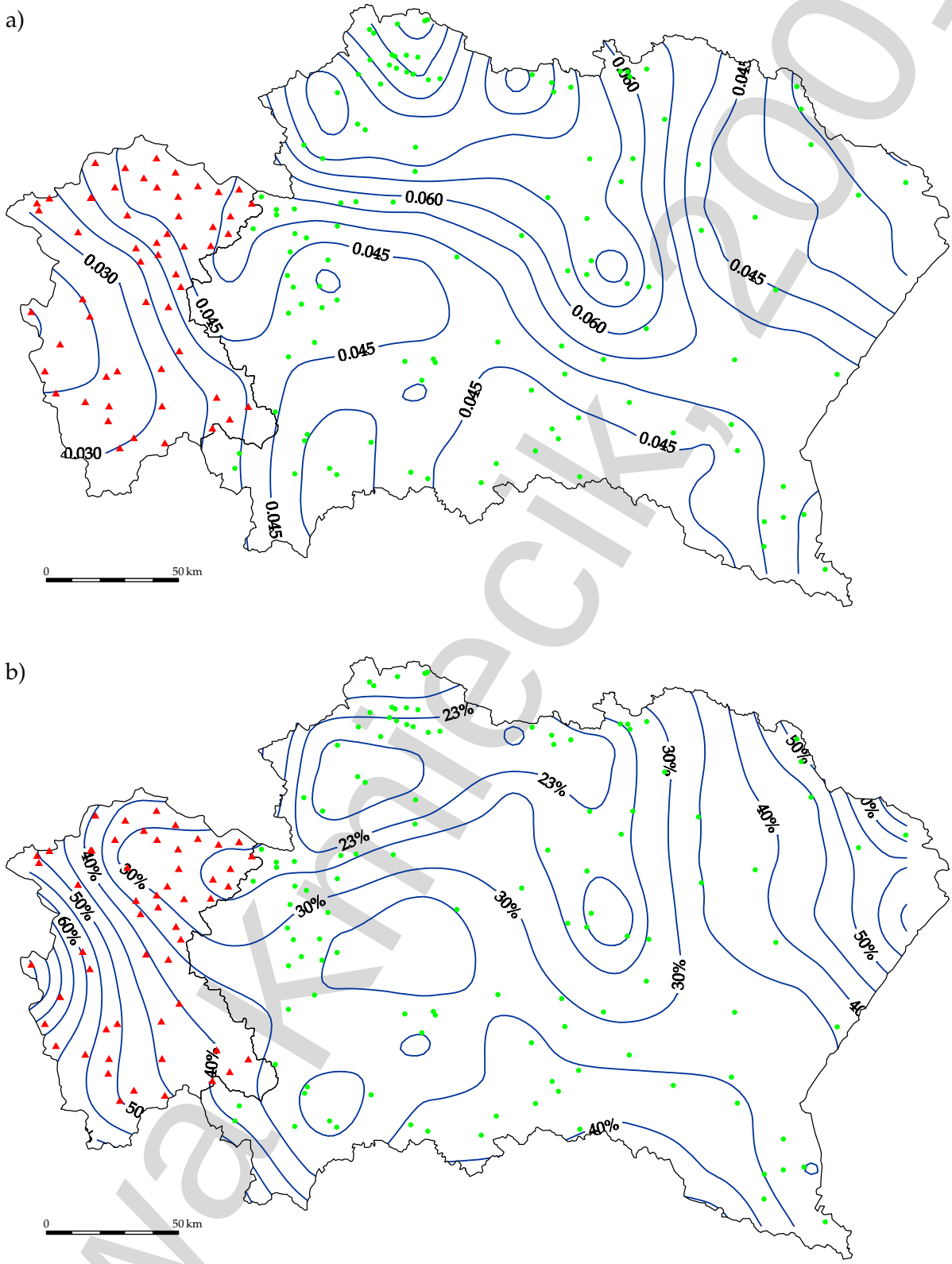
Rysunek 1.40. Mapa izoliniowa rozkładu siarczanów [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu Gaussa o parametrach $C_0 = 800$, $C = 360$, $a = 50\,000$ m. Opróbowanie: 1993 rok, seria 1



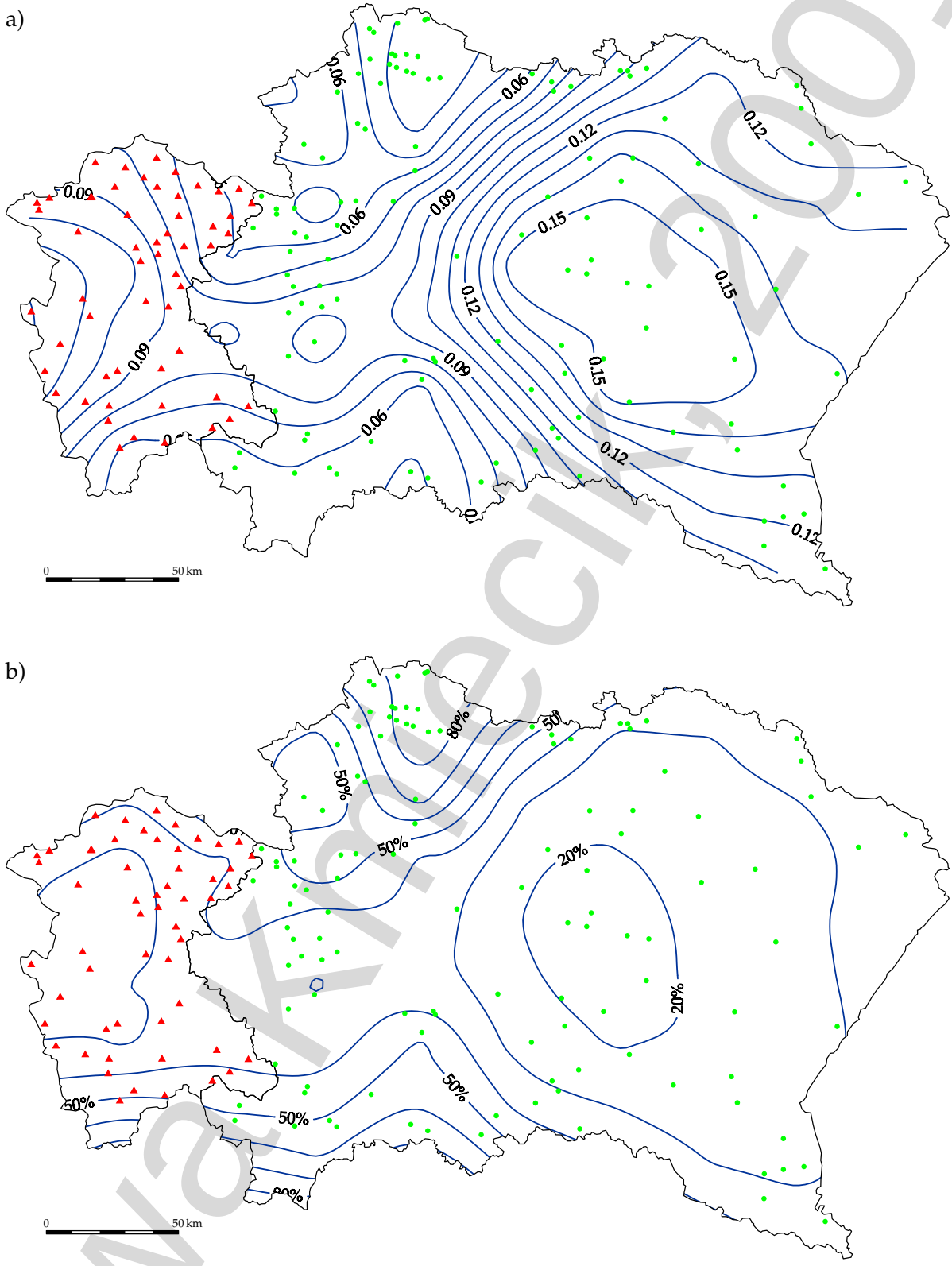
Rysunek 1.41. Mapa izoliniowa rozkładu krzemionki [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu liniowego o parametrach $C_0 = 10$, $C = 20$, $a = 40\,000$ m. Opróbowanie: 1993 rok, seria 1



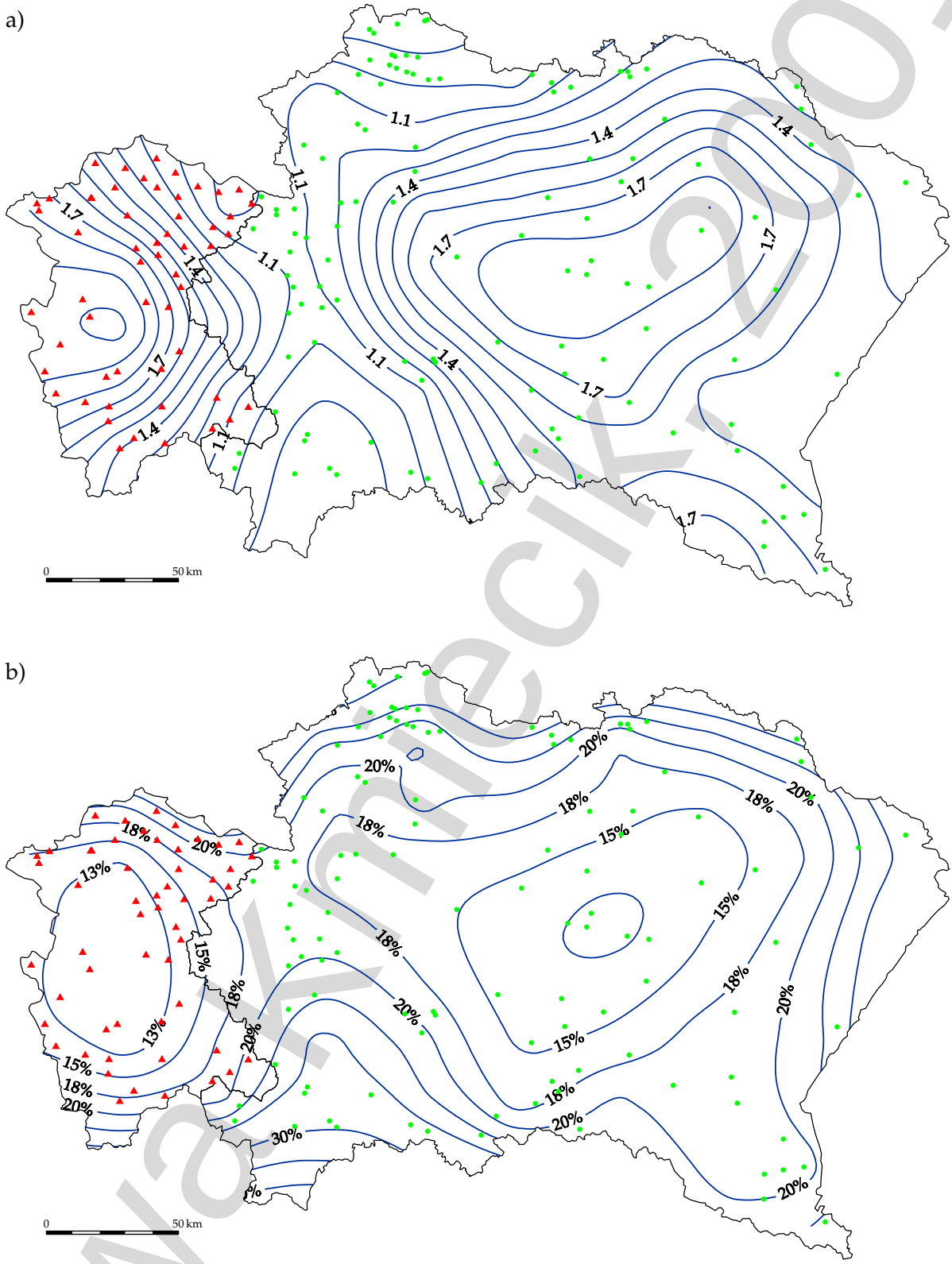
Rysunek 1.42. Mapa izoliniowa rozkładu fluorków [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą kriging-gu blokowego wg modelu Gaussa o parametrach $C_0 = 0.007$, $C = 0.02$, $a = 200\,000$ m. Opróbowanie: 1993 rok, seria 1



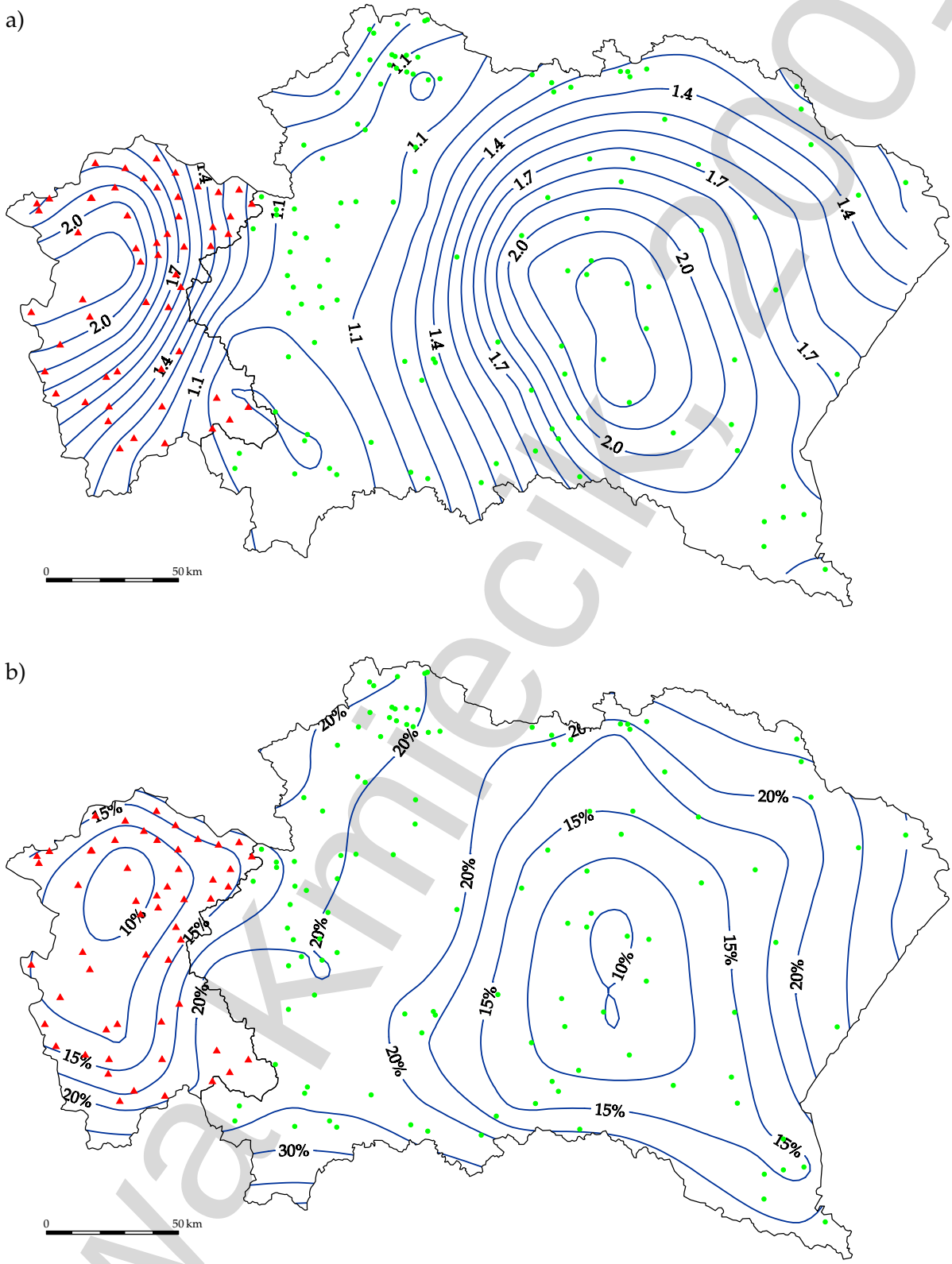
Rysunek 1.43. Mapa izoliniowa rozkładu cynku [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą kriggingu blokowego wg modelu sferycznego o parametrach $C_0 = 0.0016$, $C = 0.0004$, $a = 60\,000$ m. Opróbowanie: 1993 rok, seria 1



Rysunek 1.44. Mapa izoliniowa rozkładu współczynnika absorpcji UV (A 254) w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu liniowego o parametrach $C_0 = 0.0075$, $C = 0.0015$, $a = 40\,000$ m. Opróbowanie: 1993 rok, seria 1



Rysunek 1.45. Mapa izoliniowa rozkładu rozpuszczonego węgla organicznego [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu sferycznego o parametrach $C_0 = 0.4$, $C = 0.2$, $a = 100\,000$ m. Opróbowanie: 1993 rok, seria 1



Rysunek 1.46. Mapa izoliniowa rozkładu utlenialności [mg/dm^3] w obszarze dorzecza górnej Wisły (▲ — RZGW Katowice, ● — RZGW Kraków) a) i mapa błędów oszacowania b). Oszacowanie metodą krigingu blokowego wg modelu sferycznego o parametrach $C_0 = 0.35$, $C = 0.27$, $a = 120\,000$ m. Opróbowanie: 1993 rok, seria 1

rozdział 2

Sieci neuronowe

Początki sieci neuronowych sięgają lat 40 dwudziestego wieku. Przyjmuje się że dziedzina ta zaistniała wraz z pojawieniem się historycznej pracy autorów Mccullocha i Pittsa (1943). W pracy tej po raz pierwszy matematycznie opisano komórkę nerwową i powiązano ten opis z problemem analizy danych (Tadeusiewicz, 1993).

Już wówczas stwierdzono, że najbardziej istotną cechą sieci neuronowych jest ich zdolność do przetwarzania informacji w sposób równoległy, całkowicie odmienny od szeregowej pracy tradycyjnego komputera. Bardzo szybko stwierdzono też, że główną zaletą sieci jest proces ich uczenia się, zastępujący tradycyjne programowanie.

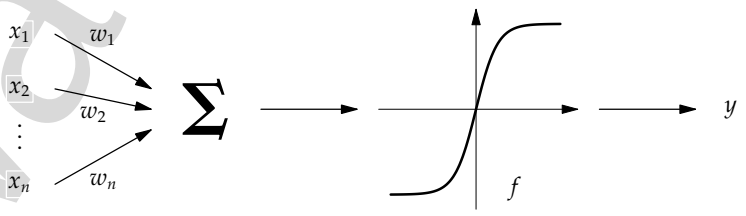
Pierwszym szeroko znanym przykładem działającej sieci neuropodobnej był Perceptron, zbudowany w 1957 roku przez Rosenblatta i Wightmana. Jego przeznaczeniem było rozpoznawanie znaków alfanumerycznych z procesem uczenia jako metodą programowania systemu (Tadeusiewicz, 1993).

W latach 70. rozwój badań nad problematyką sieci neuronowych został gwałtownie zahamowany, głównie ze względu na publikacje sugerujące ograniczony zakres zastosowań sieci, a także ze względu na fascynację osiągnięciami techniki komputerowej i rozwojem technologii układów scalonych.

Sieci neuronowe nabrały znaczenia w latach 80., gdy już na szeroką skalę pojawiły się komputery i oprogramowanie, umożliwiające ich modelowanie i rozwój technik modelowania. Sieci neuronowe okazały się narzędziem bardzo przydatnym przy realizacji wielu różnorodnych zadań.

2.1. Charakterystyka sieci neuronowych

Pierwowzorem wszelkich sieci neuronowych jest mózg ludzki(1). Rutynowo wykonuje on czynności, z którymi najszybsze komputery nie są w stanie sobie poradzić.



Rysunek 2.1. Schemat sieci neuronowej (SPSS, 1997). Objaśnienia: $x_1, \ldots, x_n$ — sygnały wejściowe; $w_1, \ldots, w_n$ — wagi połączeń; f — funkcja aktywacji neuronu; y — sygnał wyjściowy

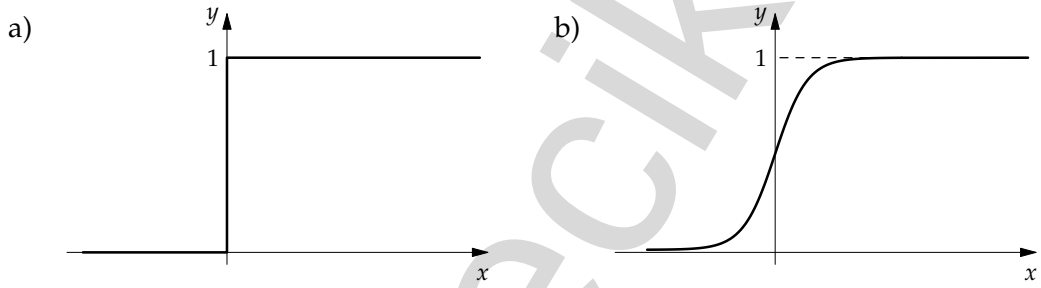
(1) Szczegóły dotyczące budowy układu nerwowego człowieka, sposobu generowania połączeń można znaleźć w literaturze medycznej oraz z zakresu sieci neuronowych, np. Tadeusiewicz 1993, 1999, 2001.

Sieć neuronowa to bardzo uproszczony model mózgu. Składa się ona z dużej liczby (od kilkuset do kilkudziesięciu tysięcy) elementów przetwarzających informację. Elementy te nazywane są neuronami, chociaż w stosunku do rzeczywistych komórek nerwowych są bardzo uproszczone.

Do neuronu dociera pewna liczba sygnałów (wartości) wejściowych. Są to albo wartości danych pierwotnych, podawanych do sieci z zewnątrz jako dane do prowadzonych w sieci obliczeń, albo sygnały pośrednie (pochodzące z wyjść innych neuronów wchodzących w skład sieci). Każda wartość wprowadzana jest do neuronu przez połączenie o pewnej sile (tzw. wadze), modyfikowanej w trakcie procesu „uczenia".

Każdy neuron posiada również pojedynczą wartość progową, określającą jak silne musi być jego pobudzenie. W neuronie obliczana jest ważona suma wejść (suma wartości sygnałów wejściowych mnożonych przez odpowiednie współczynniki wagowe) a następnie odejmowana jest od niej wartość progowa.

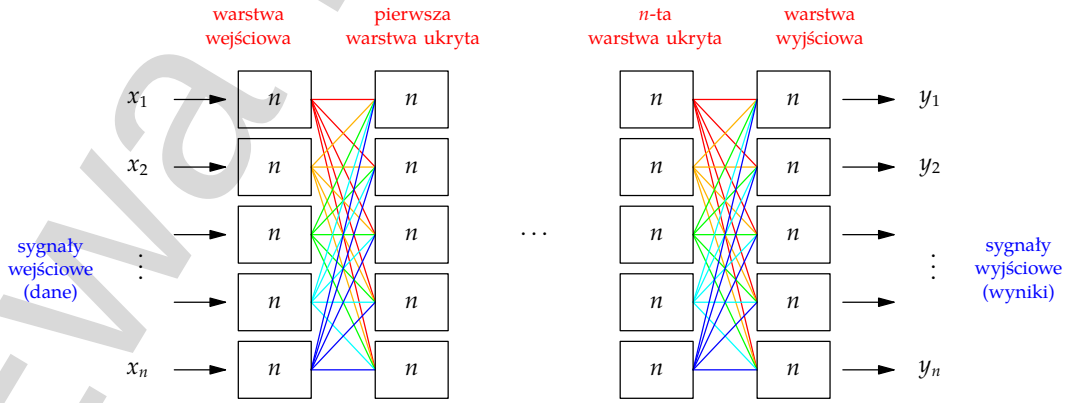
Sygnał reprezentujący łączne pobudzenie neuronu przekształcany jest przez funkcję aktywacji neuronu, a wartość obliczona przez funkcję aktywacji jest wartością wyjściową, sygnałem wyjściowym neuronu (rys. 2.1).



Rysunek 2.2. Funkcje aktywacji: a) progowa, b) sigmoid (SPSS, 1997)

Zachowanie neuronu (i całej sieci) uzależnione jest od rodzaju użytej funkcji aktywacji. Jeśli zastosowana zostanie progowa funkcja aktywacji (rys. 2.2a), to sztuczny neuron działa podobnie do neuronu biologicznego. W rzeczywistości progowa funkcja aktywacji jest bardzo rzadko stosowana w sztucznych sieciach neuronowych, ze względu na kłopoty podczas uczenia (Tadeusiewicz, 1999). Najczęściej stosowane są funkcje aktywacji dostarczające sygnałów o wartościach zmieniających się w sposób ciągły, np. sigmoid (rys. 2.2b).

Elementy, z których budowane są sieci neuronowe — neurony — charakteryzują się występowaniem wielu wejść i jednego wyjścia. W większości stosowanych modeli sieci neuronowych grupy neuronów są uporządkowane w warstwy (struktura).



Rysunek 2.3. Struktura sieci neuronowej (SPSS, 1997)

Warstwa wejściowa pobiera sygnał z otoczenia (rys. 2.3). Sygnał wyjściowy tej warstwy staje się sygnałem wejściowym dla pierwszej warstwy ukrytej. Warstwa wyjściowa pobiera sygnał wejściowy, będący sygnałem wyjściowym ostatniej warstwy ukrytej i wysyła sygnał wyjściowy do otoczenia.

Liczba neuronów w warstwie wejściowej i wyjściowej jest określona liczbą wejściowych i wyjściowych danych. W najprostszym przypadku w sieciach służących do predykcji każdej prognozowanej cesze odpowiada jeden neuron.

Wyróżnia się dwa główne typy sieci neuronowych:

- sieci jednokierunkowe (*feed-forward*);
- sieci ze sprzężeniem zwrotnym (*feedback*).

W przypadku sieci jednokierunkowych sygnał przepływa tylko w jednym kierunku — od wejść, poprzez neurony ukryte, do neuronów wyjściowych. W systemie nie ma żadnych pętli ze sprzężeniem zwrotnym. Raz nauczona (wytrenowana) sieć tego typu zawsze daje taką samą odpowiedź na dany sygnał wejściowy. Przykładem takiej sieci jest wielowarstwowy perceptron MLP, pod pojęciem tym należy rozumieć rodzinę sieci, w których uczenie odbywa się poprzez wsteczną propagację błędu przez sieć (dlatego często sieci te nazywane są sieciami wstecznej propagacji).

W sieciach ze sprzężeniem zwrotnym sygnał wyjściowy neuronu może być połączony z sygnałem wejściowym (istnieją połączenia powrotne, od późniejszych do wcześniejszych neuronów). Sygnały wyjściowe w sieciach tych zawsze zależą od poprzedniego stanu sieci. Do tej grupy zaliczana jest np. sieć Hopfielda.

Sieci jednokierunkowe zwykle składają się z kilku warstw. Każda warstwa pobiera sygnał wejściowy będący sygnałem wyjściowym warstwy poprzedniej. Sieci o strukturze sprzężenia zwrotnego (*feedback*) są bardziej złożone, niektóre z modeli mają połączenia pomiędzy neuronami wewnątrz warstwy (SPSS, 1997).

Konwencjonalne techniki komputerowe są idealne do różnego rodzaju rozwiązań liniowych, w przypadku gdy nie da się w łatwy sposób utworzyć modelu matematycznego systemu, nie zachowują się najlepiej. Klasyczne obliczenia należy wcześniej zdefiniować, zaprogramować krok po kroku; sieć neuronowa aby rozwiązać problem musi być „szkolona", sama siebie programuje w bardzo efektywny sposób (uczy się złożonych szablonów, obrazów i trendów danych).

Sieci neuronowe mogą mieć zastosowanie w bardzo wielu różniących się od siebie dziedzinach, od finansów, poprzez medycynę, zastosowania inżynierskie, geologię, geodezję czy fizyka. Spośród bardzo wielu obszarów wykorzystania sieci neuronowych, opisanych w literaturze, można wymienić m.in. (Tadeusiewicz, 1993; Hippe, 2000):

- diagnostykę układów elektronicznych;
- badania psychiatryczne;
- prognozy giełdowe;
- prognozowanie sprzedaży;
- poszukiwania ropy naftowej;
- prognozy cen;
- analizy badań medycznych;
- analizę spektralną;
- różnego rodzaju optymalizacje;
- rozpoznawanie obrazów;
- robotykę, automatykę, teorię sterowania;
- sterowanie procesów przemysłowych.

W dziedzinie geologii jak do tej pory wykorzystywano jedynie metody rozpoznawania obrazów, m.in. do rozpoznawania złóż (Kotlarczyk et al., 1995; Kalabiński i Mastej, 1995; Blaschke, 1995; Waksmundzki, 1995; Kotlarczyk et al., 1997; Kotlarczyk et al., 1999).

Metoda rozpoznawania obrazów (algorytm *k*-tego najbliższego sąsiada) została zastosowana przez Zamorską (1999) do prognozowania wybranych właściwości wód powierzchniowych.

W ubiegłym roku ukazała się monografia (Gruszczyński, 2000) prezentująca sposób wykorzystania klasyfikatorów neuronowych do symulacji skutków przekształceń gleb na terenach górniczych.

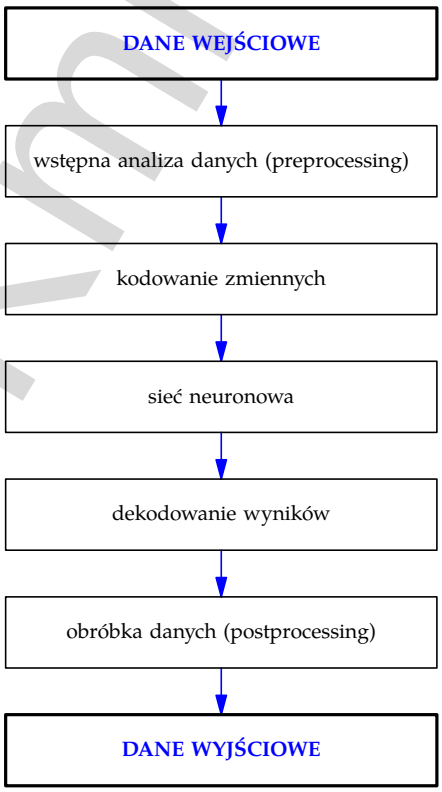
Sieci neuronowe są szeroko wykorzystywane do analizy szeregów czasowych (Lula, 2001; Zhu Mu-Lan et al., 1994; Zhang et al., 1994; Lachtermacher et al., 1994; Nabagło 1994; Świercz, 1994; Petridis, Kehagias 1998). Za ich pomocą rozwiązywane są zagadnienia prognozy nieliniowych sygnałów losowych, hydrologicznych szeregów czasowych, dziennego zapotrzebowania na wodę, czy odpływu wód ze zbiornika.

Prowadzone są również próby wykorzystania sieci neuronowych do prognozowania zmian jakości wód w układzie czasowym (Szczepańska, Kmiecik, 2000; Kmiecik 2000; Szczepańska, Kmiecik, 2001).

Cytując za Tadeusiewiczem (1999): „...sieci neuronowe mogą być zastosowane z dużym prawdopodobieństwem sukcesu wszędzie tam, gdzie pojawiają się problemy związane z tworzeniem modeli matematycznych pozwalających automatycznie (w wyniku tzw. procesu uczenia) odwzorować w komputerze różne złożone zależności pomiędzy pewnymi sygnałami wejściowymi a wybranymi sygnałami wyjściowymi".

2.2. Zastosowanie sieci neuronowych do predykcji i klasyfikacji

Zagadnienie **predykcji** polega na przypisaniu badanemu przypadkowi określonej wartości liczbowej, np. prognozowanie stężeń wskaźników fizyko-chemicznych wód w danym punkcie monitoringowym na podstawie współrzędnych tego punktu (SPSS, 1997).



Rysunek 2.4. Schemat analizy danych za pomocą sieci neuronowych (SPSS, 1997)

Pod pojęciem **klasyfikacji** należy rozumieć przypisanie jednostki do jednej z kilku kategorii (klas), np. przypisanie danego punktu monitoringowego do określonej klasy zagrożenia wód, czy obszaru o określonym zagospodarowaniu terenu na podstawie wyników oznaczeń wskaźników fizyko-chemicznych wód w tym punkcie (SPSS, 1997).

Samo zbudowanie sieci neuronowej jest jednym z wielu etapów tworzenia modelu systemu do rozwiązywania zagadnień predykcji i klasyfikacji. Sieć należycie będzie spełniała swoją rolę (poprawne prognozy, itp.), wówczas gdy dane, które będą modelowane zostaną poddane pewnej obróbce (rys. 2.4).

Pierwszym krokiem w modelowaniu danych za pomocą sieci neuronowych jest przygotowanie danych do analizy (*preprocessing*), obejmujące wiele technik analizy danych w celu selekcji tych, które zostaną użyte w modelu sieci.

Następnie należy upewnić się, że dane są zakodowane w formacie kompatybilnym z modelem sieci — etap ten obejmuje np. kodowanie zmiennych typu jakościowego (zmiennych kategorycznych) czy normalizację zmiennych ciągłych (typu ilościowego).

Po doprowadzeniu danych do postaci dogodnej do prezentacji, sieć neuronowa uczy się odwzorowania reprezentowanego przez dane, by po zbudowaniu modelu prognozować wartość docelową dla tych danych, które nie były wykorzystane w trakcie uczenia.

Odpowiedź sieci musi zostać następnie zdekodowana, a w niektórych przypadkach, uzależnionych od natury badanego przypadku podlega procesowi dalszej obróbki, tzw. *postprocessingu*.

2.2.1. Przygotowanie danych do analizy

Przed utworzeniem modelu należy dokonać wstępnej weryfikacji dostępnej bazy danych, by stwierdzić, czy zawiera ona przydatne dla modelu informacje.

Wstępna analiza danych jest konieczna w celu zapewnienia odpowiednio wysokiej dokładności modelu. Błędem w analizie wykonywanej za pomocą sieci neuronowych jest „ładowanie" do systemu wszystkich danych, jakimi się dysponuje, zakładając że sieć oddzieli sobie dane poprawne od błędnych.

W celu ułatwienia zadania predykcji, przed wprowadzeniem danych do modelu, należy dokonywać odpowiednich ich transformacji:

- usuwać zmienne ze stałymi wartościami danych (lub np. zmienne, w których istotnie przeważa jedna kategoria danych);
- transformować zmienne o rozkładzie asymetrycznym (dla sieci neuronowej idealnym byłby rozkład jednostajny), poddawać je np. operacji logarytmowania;
- stosować normalizację zmiennych, co spowoduje uniknięcie sytuacji, w której zmienna o większej wartości średniej i większej wariancji ma większy wpływ na odpowiedź sieci;
- usuwać obserwacje obarczone błędami grubymi — obserwacje nietypowe, które zniekształcają dane (Luszniewicz, Słaby, 1997; SPSS, 1997; Szczepańska, Kmiecik, 1998).

Dane w formacie znakowym, czy w formacie daty, przed wprowadzeniem do modelu sieci neuronowej muszą być przekonwertowane na format numeryczny, np. datę można przekodować na format numeryczny postaci YY.YY (na przykład: format 92.5 oznacza datę 30/06/92). Innym, lepszym sposobem jest wyrażenie daty np. w postaci liczby dni, jaka minęła od pewnej daty odniesienia (np. liczba dni od początku eksperymentu).

W celu nauczenia i testowania modelu potrzebna jest odpowiednia liczba danych tzw. „historycznych", zawierających informacje o zachowaniu się systemu, który ma być modelowany. Ważne jest by dane niosły informacje dotyczące całego, pełnego zakresu modelowanych zachowań, aby ustrzec się przypadków, których sieć się „nie nauczyła" — prognozy wówczas nie będą rzetelne.

Zbiór danych wejściowych dzielony jest na trzy podzbiory:

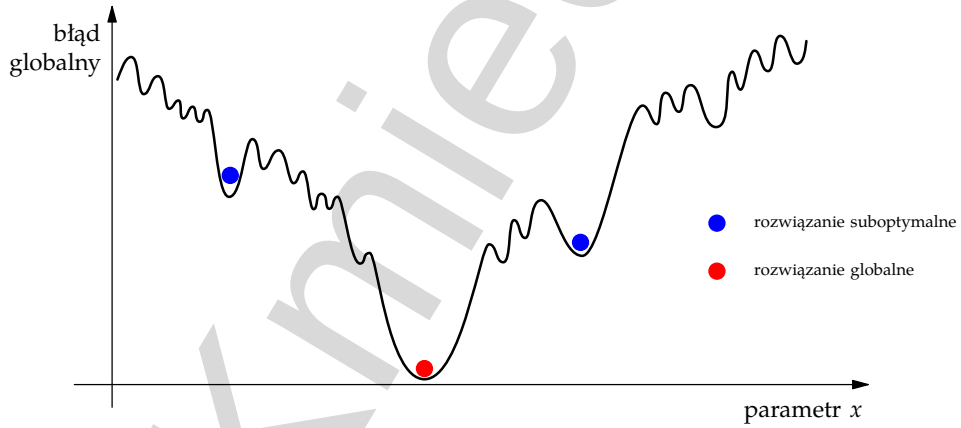
- **treningowy** — inaczej uczący, prezentowany sieci w trakcie uczenia i służący do modyfikacji parametrów (zbiór, na którym sieć uczy się odwzorowania danych);
- **walidacyjny** — inaczej weryfikacyjny, zbiór pozwalający na monitorowanie procesu uczenia sieci;
- **testowy** — służący do stwierdzenia poprawności zbudowanego modelu.

Rodzaj problemu jaki sieć ma modelować ma wpływ na wybór techniki podziału danych na podzbiory: treningowy, walidacyjny i testowy. Jeśli problem dotyczy predykcji statycznej lub jest zadaniem polegającym na klasyfikacji, wówczas obserwacje do zbiorów testowego, treningowego i walidacyjnego powinny być wybierano losowo, tak by uniknąć wprowadzenia czasowych zależności do modelu (gdyż np. wyniki uzyskane w latach 1990–1993 mogą mieć zupełnie inne cechy niż wyniki z lat 1999–2000).

2.2.2. Budowa modelu sieci neuronowej

Istnieje wiele różnych rodzajów sieci neuronowych, z których każdy ma własną charakterystykę. To, jaki rodzaj sieci neuronowej jest najbardziej przydatny do danego modelu, zależy od wielu różnych czynników.

Celem w analizie sieci neuronowych jest znalezienie rozwiązania globalnego (rys. 2.5) w obszarze, który zawiera kilka suboptymalnych rozwiązań. Rozwiązanie globalne daje najmniejszy możliwy błąd — najmniejszy, gdyż w większości przypadków nie można znaleźć perfekcyjnego modelu rozwiązania globalnego, tak by błąd wynosił zero.



Rysunek 2.5. Ilustracja obszaru błędu (SPSS, 1997)

Większość sieci neuronowych uczy się właściwego odwzorowania „wejścia" na „wyjście" poprzez minimalizację błędu pomiędzy wartością prognozowaną a wartością prawdziwą (docelową). W przypadku złożonych problemów może istnieć kilka rozwiązań suboptymalnych. Trudność polega na takim skonfigurowaniu sieci, by proces uczenia nie zatrzymał się na jednym z takich suboptymalnych rozwiązań.

Topologie sieci zmieniają się od prostych po bardziej złożone. Mogą być wykorzystane do rozwiązywania następujących grup problemów:

- klasyfikacji — przypisanie nowej jednostki do jednej z N grup;
- predykcji — oszacowanie wartości dla nowej obserwacji;
- prognozowania szeregów czasowych — zadania wykorzystujące informacje uporządkowane w czasie (predykcja w oparciu o informacje historyczne tej samej cechy);
- segmentacji danych — podział dużych baz danych na klastry w oparciu o podobieństwo obserwacji.

Pierwsze trzy grupy problemów mogą być rozwiązane za pomocą klasy modeli sieci neuronowych znanych jako modele nadzorowane (z nauczycielem, *supervised*). W przypadku tych modeli musi być podana zmienna docelowa. Algorytm uczący działa jak mechanizm nauczyciela, modyfikując wagi sieci, tak że model uczy się odwzorowania danych wejściowych na wartości docelowe.

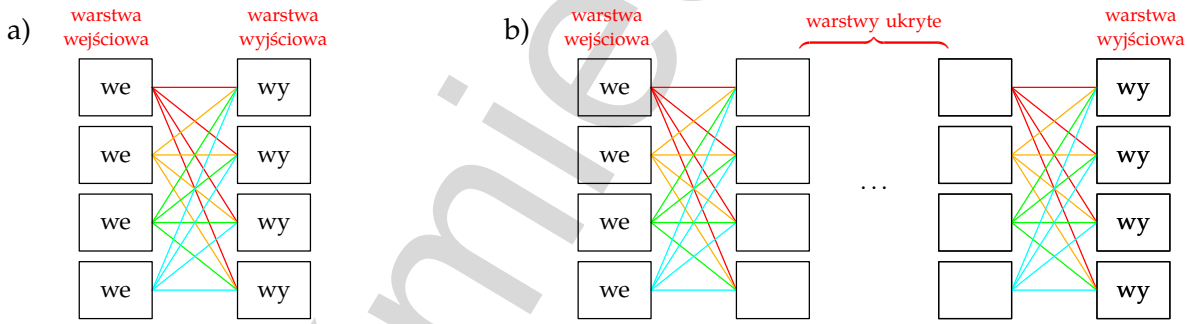
Problem segmentacji danych wymaga zastosowania klasy modeli sieci neuronowych nienadzorowanych (bez nauczyciela, *unsupervised*). Takie struktury sieci neuronowych nie potrzebują zmiennej docelowej, „uczą się" na podstawie korelacji między danymi. Najbardziej popularnym przykładem sieci działającej „bez nauczyciela" jest sieć Kohonena (SPSS, 1997).

Aktualnie jest dostępnych wiele modeli sieci nadzorowanych, chociaż tak naprawdę są one wariantami, czy odmianami ograniczonej liczby modeli. Bardziej szczegółowo przedstawione zostaną wykorzystywane w niniejszej pracy modele sieci „z nauczycielem": Multi-Layer Perceptron (wielowarstwowy perceptron), Radial Basis Function (radialna funkcja bazowa) i sieć Bayesa.

2.2.3. Perceptron wielowarstwowy (Multi-Layer Perceptron)

Model Multi-Layer Perceptron (MLP) powstał w roku 1980 i był pierwszym modelem, który mógł przetwarzać dane nieliniowe. Jest narzędziem do modelowania i prognozowania, może być zatem wykorzystany do rozwiązywania problemów klasyfikacji i predykcji.

Model Multi-Layer Perceptron jest siecią neuronową opartą na oryginalnym modelu prostego perceptronu, z dodatkowymi warstwami neuronów ukrytych pomiędzy warstwą wejściową i wyjściową, pozwala zatem na dokładne odzwierciedlenie nieliniowości danych.



Rysunek 2.6. Model zwykłego (pojedynczego) perceptronu (a) i perceptronu wielowarstwowego (b) (SPSS, 1997)

Zwykły perceptron składa się z warstwy wejściowej i wyjściowej, bez warstw ukrytych. Każdy neuron w warstwie wejściowej jest połączony z każdym neuronem w wartwie wyjściowej. W trakcie uczenia sieci dopasowywane są „odległości" pomiędzy neuronami (rys. 2.6a).

Wynik dla neuronu w perceptronie jest iloczynem wartości wprowadzonych na wejściu i odpowiednich wag. Uzyskując obraz wejściowy, perceptron tworzy zestaw wartości wyjściowych, który zależy od obrazu wejściowego i wartości połączeń.

Zwykłe perceptrony mogą rozwiązywać zagadnienia liniowo separowalne, do zagadnień nieseparowalnych liniowo należy wykorzystać perceptron wielowarstwowy (*Multi-Layer Perceptron*).

Perceptron wielowarstwowy (MLP) różni się od pojedynczego perceptronu istnieniem warstw neuronów ukrytych oraz faktem wykorzystania funkcji aktywacji do modyfikacji wejścia neuronu (rys. 2.6b). Aktywacja warstw ukrytej i wyjściowej odbywa się w taki sam sposób jak w przypadku zwykłych perceptronów, ale funkcja transferu jest wygładzoną funkcją nieliniową, zwykle funkcją sigmoidalną (z tego względu, że algorytm wymaga takiej funkcji odpowiedzi, która ma ciągłą pierwszą pochodną).

Proces uczenia przebiega w następujący sposób: najpierw są inicjalizowane wagi i wartości progowe (*bias*), najczęściej jako małe liczby losowe. Obraz treningowy jest przypisany wówczas do jednostek wejściowych i obliczane są aktywacje neuronów w pierwszej warstwie ukrytej. Wyniki dla tych neuronów przez funkcję transferu przesyłane są do neuronów w kolejnej warstwie. Proces takiego przesyłania „w przód" jest powtarzany aż do momentu kiedy otrzymany zostanie wynik w warstwie wyjściowej (SPSS, 1997).

Następnie mierzona jest różnica pomiędzy wartością uzyskaną jako wynik, a wartością prawdziwą (docelową) i „długości" połączeń w sieci są zmieniane, tak by wyniki uzyskane na wyjściu były jak najbliższe wartościom docelowym. Osiągane to jest w przebiegu powrotnym, „wstecznym" — od neuronów wyjściowych do wejściowych.

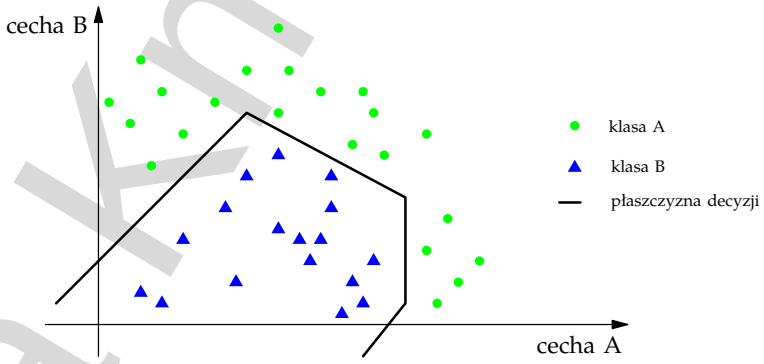
Reguła uczenia (*learning rule*) służąca do zmiany połączeń jest bardzo prosta. Jeśli wynik uzyskany na wyjściu jest poprawny, połączenia od neuronów wyjściowych do wejściowych nie są zmieniane. Jeśli wynik uzyskany na wyjściu jest większy niż wartość docelowa, połączenia pomiędzy danym neuronem wyjściowym a neuronami wejściowymi są zmniejszane, i odwrotnie.

W takim przypadku istnieje jednak ryzyko osiągnięcia tzw. minimum lokalnego (rys. 2.5), zatem aby znaleźć optymalną wartość, należy uruchamiać algorytm z różnymi wartościami początkowymi, oraz zmieniać parametry uczenia: szybkość (*learning rate*) i bezwładność (*momentum*). Parametry te powinny przyjmować wartości z przedziału 0.1–0.9, mniejszym niebezpieczeństwem jest tu przyjęcie za dużych niż za małych współczynników (Tadeusiewicz, Mikrut, 1994).

Algorytm uczący modyfikuje wagi połączone z każdym przetwarzanym elementem, tak że system minimalizuje błąd pomiędzy wartością docelową a aktualną odpowiedzią sieci.

Aby przeprowadzić odwzorowanie nieliniowe potrzebna jest przynajmniej jedna ukryta warstwa neuronów. Liczba neuronów w sieci powinna zależeć od złożoności modelowanego systemu. Topologia wielowarstwowa jest poprawna, jednak formalnie nie ma potrzeby rozbudowywać sieci ponad jedną warstwę ukrytą. W praktyce, nie ma żadnych korzyści z zastosowania więcej niż jednej warstwy ukrytej, a nawet wręcz przeciwnie (Tadeusiewicz, 1993).

Na rysunku 2.7 przedstawiono schematycznie problem klasyfikacji za pomocą sieci MLP, w przypadku gdy są dwie cechy wejściowe i dwie klasy danych wyjściowych.



Rysunek 2.7. Płaszczyzna decyzji w sieci typu MLP — problem klasyfikacji z dwiema cechami wejściowymi i dwiema klasami danych wyjściowych (SPSS, 1997)

Do zalet topologii MLP należy zaliczyć:

- możliwość wykorzystania do szerokiego zakresu problemów;
- zdolność do interpolacji i uogólniania;
- jeśli dane nie są pogrupowane (sklastrowane) lecz rozrzucone równomiernie, sieć MLP klasyfikuje je do ekstremalnych obszarów.

Wadą sieci Multi-Layer Perceptron jest fakt, że długo się uczy i nie zawsze osiąga optymalne rozwiązanie.

Przy opracowywaniu modelu MLP należy zwrócić szczególną uwagę na dwa parametry:

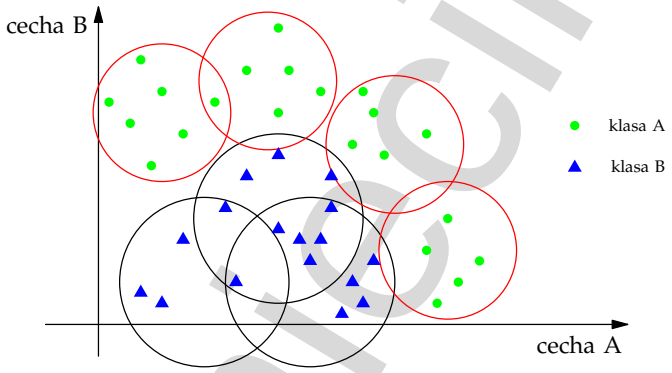
- liczbę warstw ukrytych;
- algorytm uczący.

Im więcej jest neuronów ukrytych w modelu, tym bardziej skomplikowaną funkcją będzie opisany system. Z drugiej strony, jeśli będzie za mało neuronów ukrytych, sieć nie znajdzie rozwiązania ogólnego, i możemy mieć do czynienia z efektem przeuczenia (*overtraining*). Dla każdego problemu istnieje optymalna liczba neuronów ukrytych, zależna od specyfiki zagadnienia.

Wybór algorytmu uczącego jest z kolei kompromisem pomiędzy czasem, jaki zajmuje znalezienie rozwiązania globalnego, a czasem poświęconym na obliczanie wag połączeń (SPSS, 1997). Budowanie modeli sieci neuronowych wymaga więc pewnego doświadczenia.

2.2.4. Radialna funkcja bazowa (Radial Basis Function)

Inną siecią z grupy sieci nadzorowanych jest sieć z radialną funkcją bazową RBF (*Radial Basis Function*). Ta struktura nie konstruuje płaszczyzny decyzji w przestrzeni danych wejściowych, tak jak sieć MLP — dane są klastrowane przez kilka funkcji bazowych (rys. 2.8).



Rysunek 2.8. Płaszczyzna decyzji w sieci typu RBF — problem klasyfikacji z dwiema cechami wejściowymi i dwiema klasami danych wyjściowych (SPSS, 1997)

Jeśli punkt danych leży w obszarze aktywacji danej funkcji bazowej, wówczas węzeł odpowiadający tej funkcji reaguje najmocniej (najostrzej).

Należy podkreślić, że:

- w przypadku sieci RBF proces uczenia się przebiega szybciej niż w przypadku MLP;
- sieć RBF czytelniej modeluje dane zgrupowane lokalnie niż sieć MLP.

Wadą sieci RBF jest gorsza zdolność do prezentacji ogólnych, globalnych cech danych oraz problemy z określeniem optymalnego położenia centrów funkcji radialnych.

Przy projektowaniu sieci RBF należy brać pod uwagę następujące parametry:

- liczbę centrów wymaganą do dokładnego modelowania danych;
- pozycjonowanie centrów;
- rodzaj funkcji radialnej.

Liczba centrów jest silnie uzależniona od złożoności problemu, zbyt mała ich liczba powoduje uzyskiwanie błędnych prognoz, z kolei zbyt dużo centrów daje efekt tzw. nadmiernego dopasowania (*over-fitting*), i błędnych uogólnień. Pozycjonowanie centrów zależy od sposobu, w jaki są inicjowane wagi przypisane do neuronów. Najczęściej przypisanie wag następuje losowo. Centra także wybierane są losowo ze zbioru treningowego.

Kształt nieliniowej funkcji bazowej określa odpowiedź neuronu na nowy, nieznany punkt. Przy projektowaniu sieci neuronowej należy eksperymentalnie dobrać rodzaj funkcji bazowej,

gdyż to, która funkcja jest najlepsza w danym przypadku zależy od rozkładu klas w przestrzeni cech wejściowych.

Jeśli w trakcie testowania systemu okaże się, że uzyskiwane wyniki nie są zadowalające, należy próbować dzielić dane na podgrupy, gdyż badane zmienne mogą być skorelowane w podgrupach (SPSS, 1997).

2.2.5. Sieć Bayesa (Bayesian Network Tool)

Sieć Bayesa ma strukturę podobną do modelu perceptronu wielowarstwowego, jednak do utworzenia modelu uogólnionego nie potrzebuje zbioru walidacyjnego. Może być zatem stosowana w przypadku ograniczonej liczby danych.

W sieci MLP, algorytm uczący dokonuje zmian wag połączeń między neuronami w celu zminimalizowania błędu. Ponieważ zbiór treningowy jest skończony, istnieje ryzyko, że sieć nauczy się też „szumu". W przypadku sieci Bayesa, wpływ szumu jest ograniczony, poprzez dodanie dodatkowego składnika do wyrażenia opisującego błąd.

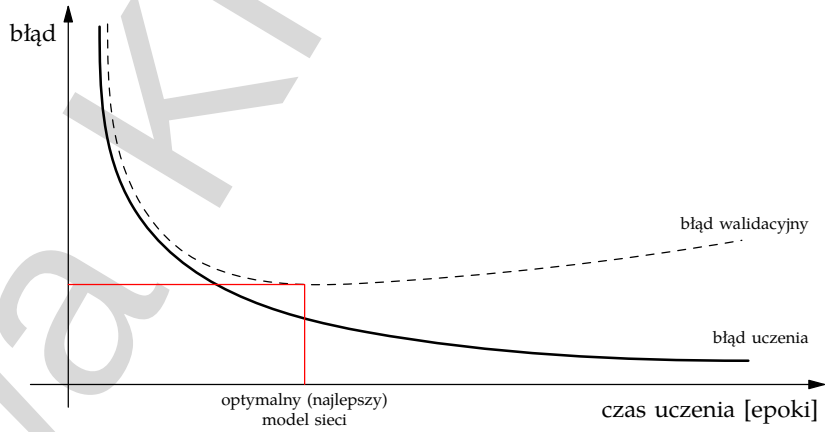
W przeciwieństwie do sieci MLP, sieć Bayesa dokonuje automatycznie normalizacji danych. W sieci tej może być jedna lub dwie warstwy ukryte neuronów (SPSS, 1997).

2.2.6. Walidacja modelu sieci neuronowej

Celem procesu uczenia sieci neuronowej jest minimalizacja błędu pomiędzy odpowiedzią systemu (wynikiem) a wartością docelową (prawdziwą). Osiągnięcie błędu zerowego oznacza, że model perfekcyjnie nauczył się charakterystyki danych ze zbioru treningowego.

Ponieważ dane treningowe zawsze zawierają pewien „szum", należy sądzić, że model sieci neuronowej nauczył się też charakterystyki szumu. Szum, z definicji, jest charakterystyką nieprognozowalną, zatem fakt, że sieć „uczy się" tego szumu może spowodować tzw. przeuczenie sieci (*overtraining*). Aby tego uniknąć, należy utworzyć zbiór danych do walidacji procesu uczenia.

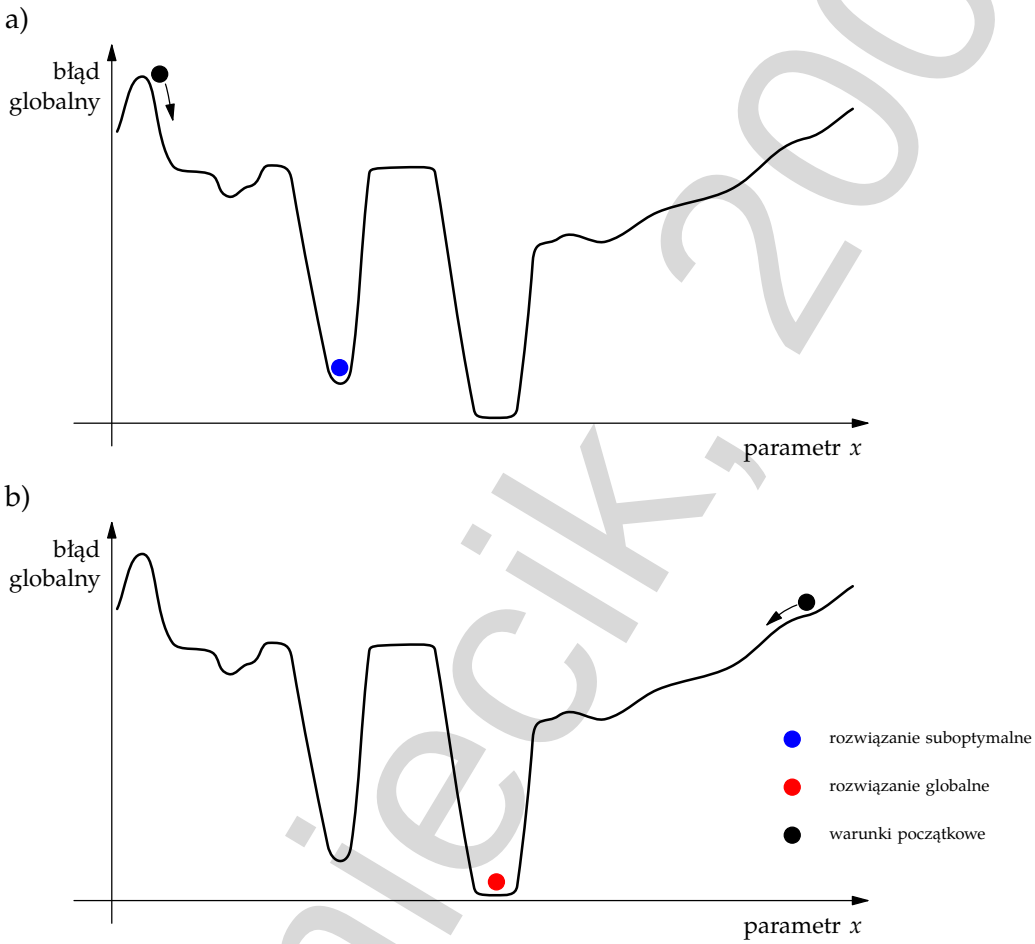
Zbiór walidacyjny zawiera mały procent danych wejściowych, nie wykorzystany przy budowaniu modelu. Dane z tego zbioru służą do monitorowania zachowania się systemu w trakcie procesu uczenia. Monitorowanie odbywa się poprzez pomiar błędu na danych walidacyjnych, w różnych odstępach czasu, w trakcie procesu uczenia (rys. 2.9).



Rysunek 2.9. Walidacja modelu sieci neuronowej (SPSS, 1997)

W początkowych fazach procesu uczenia błędy: uczenia i walidacyjny, pozostają w stałym stosunku, w momencie gdy system zaczyna uczyć się z danych treningowych charakterystyki szumu, gradient błędu walidacji maleje a na końcu wzrasta. Optymalny model sieci to model z najmniejszym błędem walidacji.

Aby uniknąć wspomnianego wcześniej osiągnięcia przez sieć rozwiązania suboptymalnego (rys. 2.5) można stosować różne warunki początkowe (rys. 2.10).



Rysunek 2.10. Ilustracja efektu zmiany warunków początkowych (SPSS, 1997)

Celem algorytmu uczącego jest doprowadzenie systemu do stanu, w którym osiągnie on najmniejszy błąd. W przypadku przedstawionym na rysynku 2.10a jest duża szansa, że osiągnięte rozwiązanie będzie rozwiązaniem suboptymalnym.

W sytuacji na rysunku 2.10b zmienione zostały początkowe wartości wag i wynik końcowy będzie prawdopodobnie rozwiązaniem optymalnym. Poprzez zmianę warunków początkowych przeprowadzona została zgrubna eksploracja powierzchni błędu.

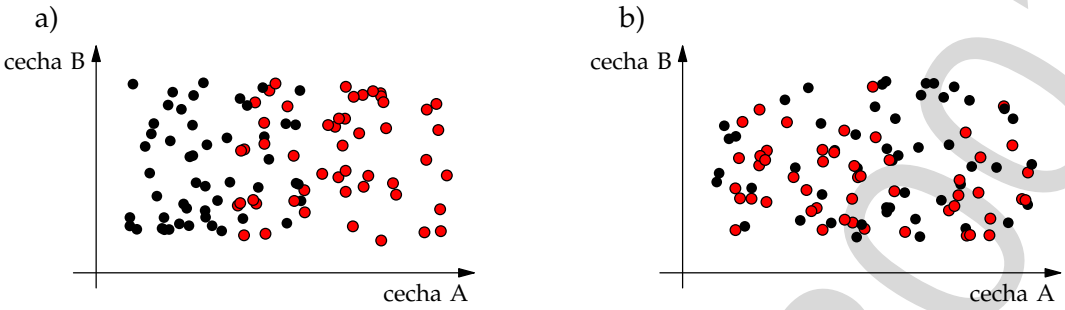
W praktyce, należy przeprowadzić przynajmniej pięć eksperymentów budowania modelu sieci neuronowej, z różnymi warunkami początkowymi. Przeciętny błąd walidacji modelu powinien być na poziomie kilku procent (SPSS, 1997).

Kolejnym sposobem na stwierdzenie poprawności zbudowanego modelu sieci neuronowej jest np. utworzenie kilku zbiorów testowych i ocena średniej poprawności modelu.

Jeśli dane były zbierane w różnym czasie, obserwacje do zbioru testowego powinny być tak dobrane by reprezentować wszystkie odcinki czasu.

W przypadku gdy liczba danych jest ograniczona, i nie ma możliwości podziału zbioru na kilka zbiorów testowych poprawność modelu można testować poprzez zróżnicowanie podziału zbioru danych na podzbiory treningowy i testowy (rys. 2.11).

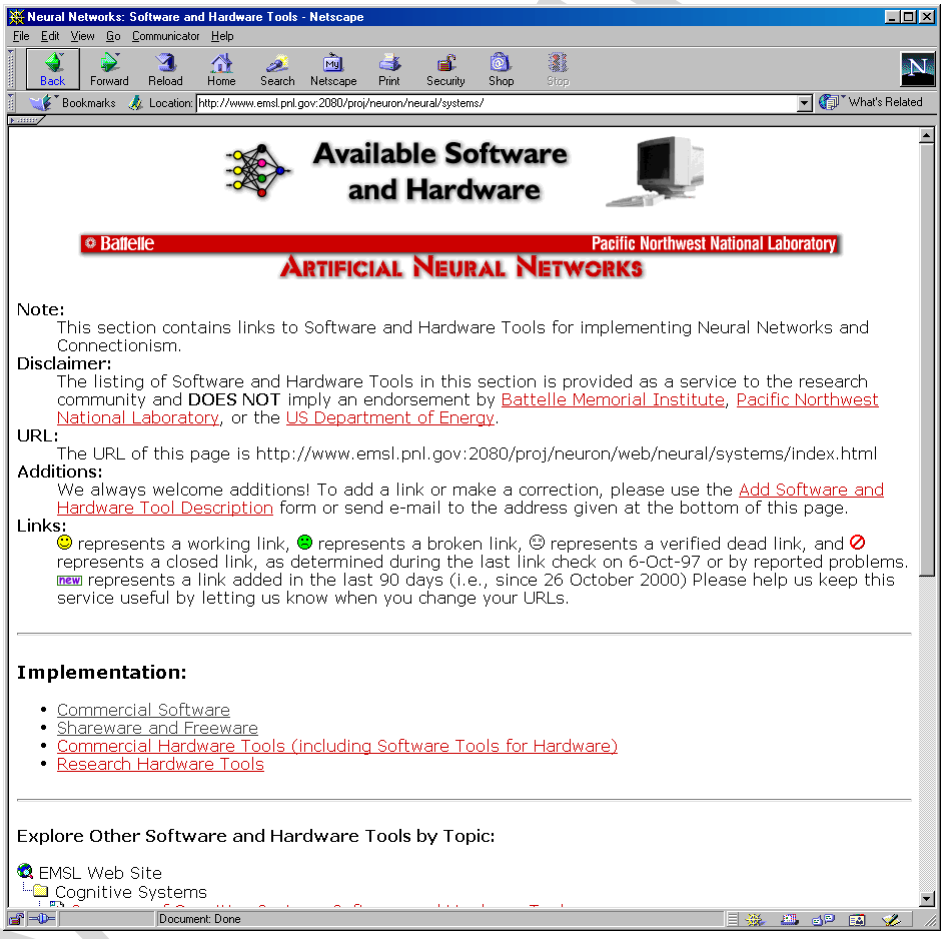
Jeżeli baza danych zawiera obserwacje zależne od czasu, obserwacje do zbiorów testowego i treningowego powinny być wybierane losowo (SPSS, 1997).



Rysunek 2.11. Sposoby wyboru zbiorów treningowego i testowego: a) błędny, b) poprawny (SPSS, 1997)

2.3. Programy komputerowe do tworzenia modeli sieci neuronowych

Większość programów komputerowych do tworzenia modeli sieci neuronowych pozwala użytkownikowi w bardzo łatwy sposób przeprowadzić analizę.



Rysunek 2.12. Strony internetowe Pacific Northwest National Laboratory. Wykaz oprogramowania i sprzętu do implementowania sieci neuronowych

Użytkownik nie musi wykazywać się znajomością skomplikowanego aparatu matematycznego, potrzebuje jedynie wiedzy dotyczącej sposobu przygotowania danych, musi dokonać wyboru rodzaju sieci neuronowej i zinterpretować uzyskane wyniki.

Czynnikiem, który bardzo często ogranicza możliwość zastosowania komputerów w analizie danych nie jest już, na szczęście, moc obliczeniowa komputera, częstotliwość taktowania procesora, rozmiar pamięci operacyjnej lub rozmiary programu, lecz niestety, koszty oprogramowania.

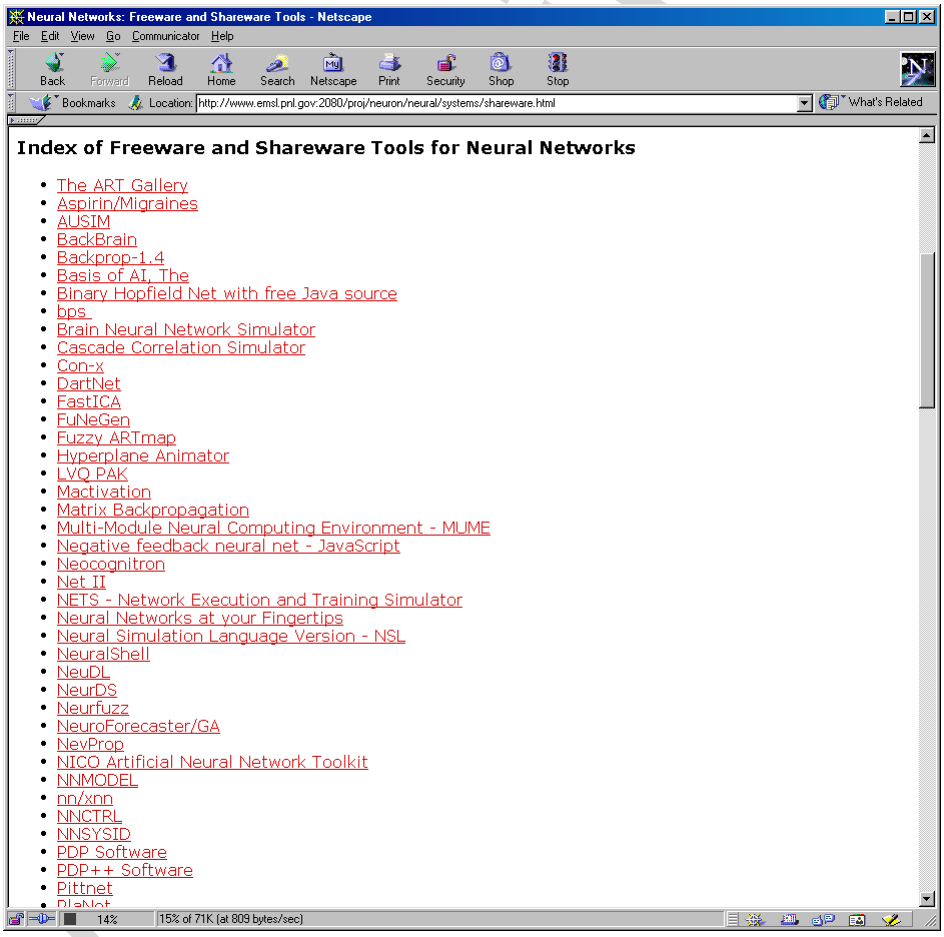
W światowej sieci Internet pod adresem:

http://www.emsl.pnl.gov:2080/proj/neuron/web/neural/systems/software

znajduje się wykaz (i krótki opis) dostępnego na wszystkie platformy systemowe oprogramowania (komercyjnego i *freeware*) dotyczącego sieci neuronowych (rys. 2.12).

Na liście tej znajdują się m.in. takie programy komercyjne, jak: Braincel, BrainMaker, Clementine, DataEngine, ECANSE — Environment for Computer Aided Neural Software Engineering, MATLAB: Neural Network Toolbox, Neural Bench, Neural Connection, NeuralWorks, NeuroLab, NeuroSolutions, SAS: Neural Network Add-On, STATISTICA: Neural Networks czy WinBrain.

Spośród grupy programów typu *freeware* można wymienić: AINET, Brain Neural Network Simulator, EasyNet, Hyperplane Animator, Mume, Neocognitron, Net II, NETS — Network Execution and Training Simulator, Neural Shell, Pittnet, Pygmalion, WinNN, Qnet2000 czy Neurooffice (rys. 2.13).



Rysunek 2.13. Strony internetowe Pacific Northwest National Laboratory. Wykaz oprogramowania freeeware i shareware do implementowania sieci neuronowych

Pod adresem

http://www.emsl.pnl.gov:2080/proj/neuron/web/neural/journals.html

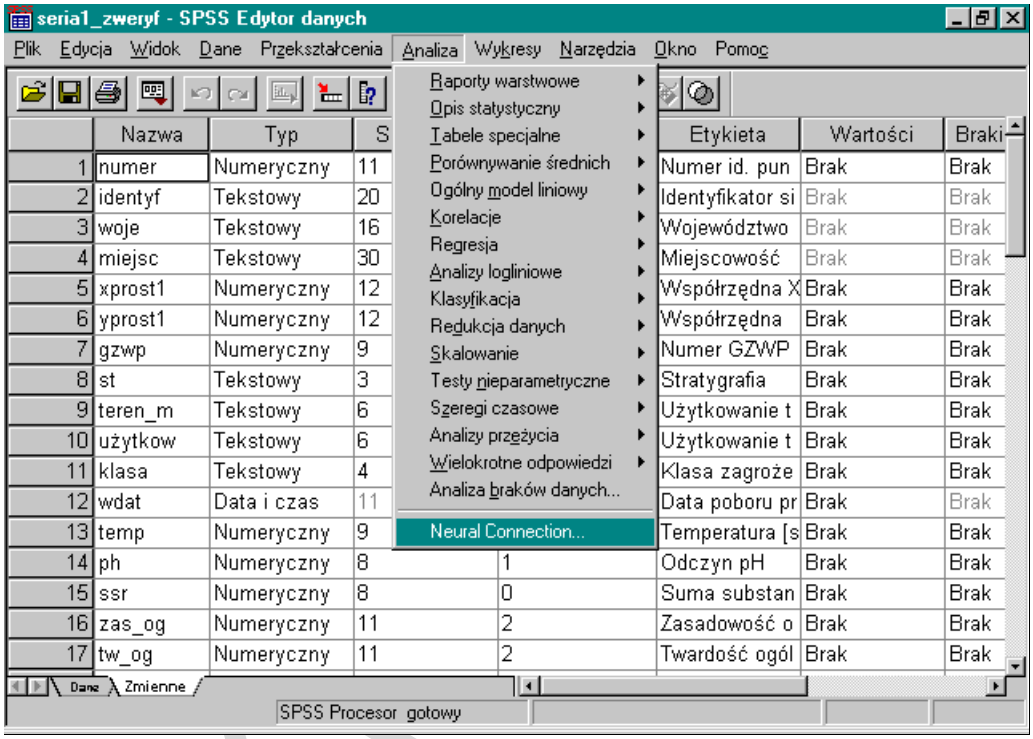
znajduje się również wykaz czasopism z różnych dziedzin zastosowań sieci neuronowych.

Próby tworzenia modeli sieci neuronowych dla potrzeb niniejszej pracy były prowadzone przy wykorzystaniu programów typu freeware: AINET, EASYNET, QNET oraz za pomocą programów komercyjnych: Clementine i Neural Connection (firmy SPSS).

Testowane programy z grupy *freeware* miały ograniczenie co do wielkości zbioru danych wejściowych, a proces uczenia na niektórych trwał nawet kilkadziesiąt godzin.

Program Clementine z kolei jest bardzo potężnym (i jednocześnie bardzo drogim) narzędziem, umożliwiającym tworzenie modeli sieci neuronowych oraz dostarczającym narzędzi do zarządzania danymi, modelowania, raportowania, tworzenia diagramów przepływu danych, działa na platformach Win98, WinNT oraz UNIX.

Ostatecznie, ze względu na kompatybilność z programami, na które Zakład Hydrogeologii i Ochrony Wód AGH posiada licencję (SPSS PL v. 10.0, QI Analyst v. 3.5 DB — bezpośrednia wymiana danych) wybrano program Neural Connection v. 2.1 (SPSS, 1997, 1999).



Rysunek 2.14. Dostęp do programu Neural Connection z poziomu programu SPSS

Po zainstalowaniu programu w systemie, w którym działa już program SPSS, opcja analizy sieci neuronowych dostępna jest wprost z menu **Analiza ▶ Neural Connection** programu SPSS (rys. 2.14). Oznacza to, że dane z otwartego w programie SPSS pliku danych automatycznie zostają wczytane jako dane wejściowe do programu Neural Connection.

2.3.1. Program Neural Connection v. 2.1

Program Neural Connection pozwala na budowanie modeli sieci neuronowych do różnego rodzaju zastosowań, i ma stosunkowo niewielkie wymagania sprzętowe i systemowe:

- komputer PC z procesorem co najmniej 386;
- system operacyjny Microsoft Windows 95, 98 lub NT 4.0;
- 8MB pamięci;
- 4MB wolnej przestrzeni na twardym dysku;
- napęd CD-ROM;

- monitor SVGA lub VGA z odpowiednią kartą graficzną;
- mysz.

Program pracuje z danymi w różnych formatach, a oparty na ikonach interfejs oraz opcja NetAgent ułatwiają, nawet początkującemu użytkownikowi, budowanie i testowanie modelu sieci, bez potrzeby zaprzęgania skomplikowanego aparatu matematycznego z zakresu teorii sieci neuronowych.

Neural Connection składa się z trzech oddzielnych grup modułów:

- interfejsu graficznego;
- modułu wykonawczego;
- narzędzi do analizy danych.

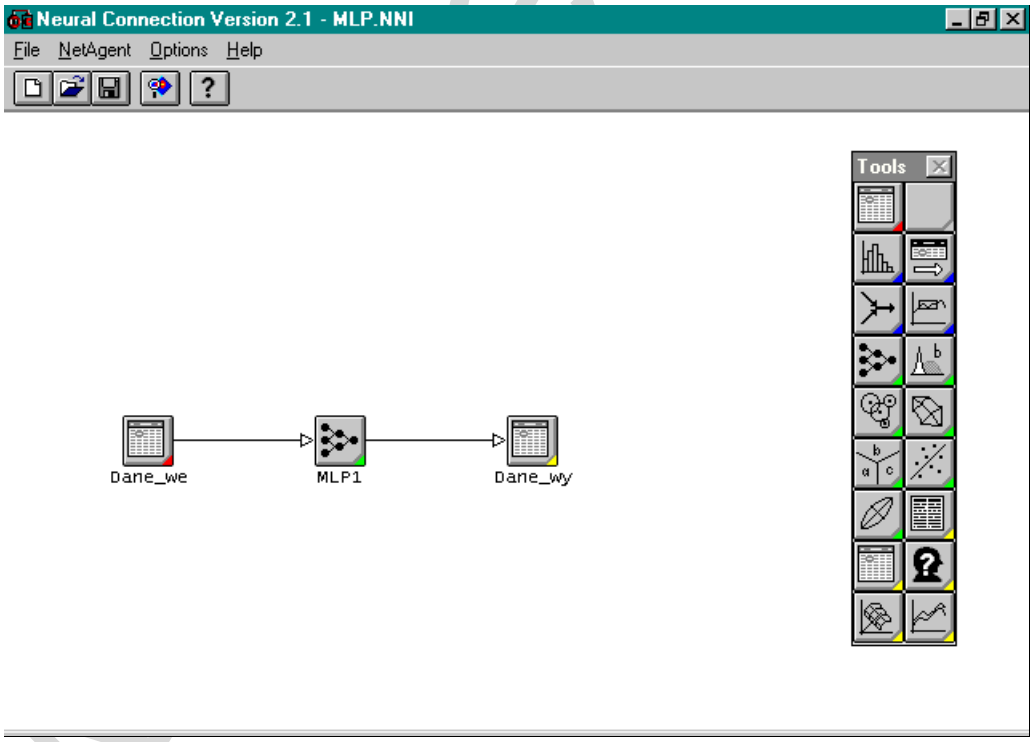
Taka budowa modułowa daje większą elastyczność niż standardowe narzędzia, umożliwia zaadaptowanie na potrzeby konkretnego, rozwiązywanego problemu.

Narzędzia, topologie i modele

Narzędzia są składnikami, z których budowana jest model. Dane, „przechodząc" przez narzędzie poddawane są określonym działaniom, prostym lub bardziej skomplikowanym. Narzędzia mogą być łączone w celu rozwiązywania określonych problemów.

Połączenia pomiędzy narzędziami definiują ścieżki i kierunki przepływu danych. Zestaw połączonych narzędzi nazywany jest topologią.

Topologia musi zawierać ścieżkę od narzędzia wprowadzającego dane (*input tool*) do narzędzia z wynikami (*output tool*), zatem najmniejsza poprawnie zbudowana topologia składa się z dwóch elementów, połączonych narzędzi typu *input* i *output*.



Rysunek 2.15. Przykład poprawnie zbudowanej topologii (SPSS, 1997)

Aby topologia była poprawna musi spełniać następujące warunki:

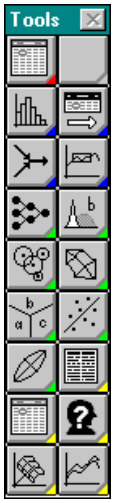
- musi rozpoczynać się od narzędzia z danymi wejściowymi (*input tool*);
- musi być zakończona narzędziem z wynikami (*output tool*);

- w topologii może być tylko jedno narzędzie z danymi wejściowymi (*input tool*);
- topologia nie może zawierać pętli ze sprzężeniem zwrotnym;
- pomiędzy narzędziem z danymi wejściowymi a narzędziem z wynikami mogą znajdować się dowolne kombinacje narzędzi.

Poprawnie zbudowana topologia (rys. 2.15) może być wykorzystana do rozwiązania problemu, zwanego inaczej modelem. Model to „urządzenie", które pomaga uzyskać istotne informacje z danych. Modele mogą być proste lub bardzo złożone, ich natura zależy od danych i od sposobu wykorzystania.

W programie Neural Connection mamy do czynienia z modelami tzw. sterowanymi danymi (*data driven*). Narzędzie nie przesyła danych do następnego narzędzia w modelu, dopóki ono nie „poprosi" o nie. To oznacza, że dane nie przepływają wzdłuż zbudowanych połączeń dopóki model nie zostanie uruchomiony.

Wybór narzędzi wykorzystanych w modelu zależy od cech analizowanego zbioru danych i od rodzaju problemu, jaki chce się rozwiązać.



Rysunek 2.16. Narzędzia do budowy topologii w programie Neural Connection

W programie Neural Connection dostępne są cztery grupy narzędzi (rys. 2.16):

- narzędzia do wprowadzania danych (*input*);
- narzędzia filtrujące (*filter*);
- narzędzia do modelowania i prognozowania (*modeling and forecasting*);
- narzędzia wynikowe (*output*).

Ikony narzędzi różnych kategorii różnią się kolorami naroży — narzędzia do wprowadzania danych mają czerwone naroża, narzędzia filtrujące — niebieskie, narzędzia do modelowania — zielone, a narzędzia wynikowe — żółte naroża. Wszystkie narzędzia są niezależne, nie ma żadnych wytycznych co do kolejności ich występowania w modelu, bardzo łatwo można je podmieniać (SPSS, 1997).

Narzędzia do wprowadzania danych (Input Tools)

Program Neural Connection ma ograniczenia dotyczące liczby danych w zbiorach treningowym i testowym: w zbiorze treningowym może znajdować się maksymalnie 750 zmiennych i 15 000 rekordów (obserwacji); zbiór testowy (roboczy) może mieć 750 zmiennych i 32 000 rekordów (obserwacji).

Program „czyta" dane różnych formatów:

- *.csv (*comma delimited*) — dane w formacie ASCII, oddzielane przecinkami;
- *.rdd (*record delimited*) — dane w formacie ASCII, kolumny oddzielane są spacjami, na końcu rekordu są znaki końca linii; liczba pól w rekordzie jest obliczana z pierwszej logicznej linii danych;

- *.nna (*field counted*) — dane są zapisane w jednym wierszu, wpisy w kolumnach są oddzielone spacjami, umieszczone są też znaki końca linii; w danych tych nie może być braków danych;
- *.txt (*fixed format*) — dane w ustalonym formacie, po wybraniu tej opcji pojawia się okno, w którym należy określić znak oddzielający kolumny i znak końca linii;
- *.sav (*SPSS files*) — program czyta pliki w formacie SPSS, od wersji 6.0, dokonując odpowiedniej konwersji danych (notacje: numeryczna, przecinkowa, z kropką, naukowa, data i czas, dolar, format użytkownika są konwertowane na format numeryczny, format znakowy jest zapisywany jako zmienna symboliczna);
- *.xls (*MS Excel 5.0*) — pliki w formacie MS Excel v. 5.0; musi to być pierwszy arkusz Excela, nie może zawierać wbudowanych obiektów czy grafiki, nie może być zabezpieczony hasłem;
- *.sys (*Systat 5.0*) — pliki w formacie programu Systat v. 5.05;
- *.* (*User defined*) — pliki w formacie użytkownika; po wybraniu tej opcji pojawia się okno, w którym trzeba podać znak oddzielający kolumny i rekordy.

W programie Neural Connection można zdefiniować pięć rodzajów danych:

- liczby całkowite (*integer*) — program rozpoznaje wszystkie liczby całkowite z zakresu od $-2\,147\,483\,647$ do $2\,147\,483\,647$, jednak lepiej nie używać liczb składających się z więcej niż 7 cyfr, gdyż program traci wówczas rozdzielczość i np. liczby 21 474 899, 21 474 900 i 21 474 901 są traktowane jako 21 474 900;
- liczby zmiennoprzecinkowe (*floating point*) — liczby z uwzględnieniem miejsc dziesiętnych, maksymalnie do 8 cyfr, opcjonalnie poprzedzone znakiem plus lub minus (podobnie jak w przypadku danych całkowitych nie należy stosować długich liczb, ze względu na utratę rozdzielczości);
- dane symboliczne (*symbolic*) — etykiety zawierające znaki alfanumeryczne są interpretowane jako ciągi znaków; program rozróżnia duże i małe litery (np. *ZMIENNA* to nie to samo co *zmienna*), dane symboliczne mogą osiągać długość do 14 znaków, wpisy dłuższe niż 14 znaków program automatycznie skraca;
- dane typu data: rok i dzień.

Program potrafi przeczytać dane zapisane w notacji naukowej, pod warunkiem poprawności zapisu, tzn.:

$$\pm\text{mantysa} \times 10^{\text{wykładnik}},$$

zatem poprawne będą zapisy: 2e3, 3.4E-2, -9e-1, 9e (taki zapis oznacza, że wykładnik jest zerowy), 2.16E2.5.

Wejściowy zbiór danych należy podzielić na trzy części:

- zbiór treningowy — na którym model „uczy się”;
- zbiór walidacyjny — wykorzystany do monitorowania zachowania systemu w trakcie uczenia sieci;
- zbiór testowy — do sprawdzenia poprawności modelu.

Proporcje, w jakich plik powinien być dzielony na poszczególne części zależą od liczby danych. Zasada jest taka, że w przypadku modelu, w którym jest N cech (zmiennych typu wejściowego) i M decyzji lub prognoz (zmiennych docelowych), co najmniej $10(M+N)$ obserwacji należy umieścić w zbiorze treningowym. Im więcej danych jest w zbiorze treningowym tym lepiej, ale im mniej danych będzie w zbiorze testowym, tym gorzej będzie sprawdzić jakość modelu, należy zatem żmudnie poszukiwać na drodze doświadczalnej pracy z programem optymalnej topologii budowanej sieci.

Narzędzie służące do wprowadzania danych posiada opcje wstępnej analizy danych (sprawdzanie zakresu analizowanych danych, usuwanie obserwacji odstających, analiza braków danych), opcje konwersji i korekcji danych.

Narzędzia wynikowe (Output Tools)

W grupie narzędzi wynikowych znajdują się mechanizmy umożliwiające uzyskiwanie wyników z modelu.

Do grupy tej należą (rys. 2.16):

- *Data Output* — uruchamia topologię i drukuje wyniki na ekranie w postaci arkusza danych, lub zapisuje wyniki do pliku (można wybrać, który ze zbiorów ma być przedstawiony: zbiór treningowy, walidacyjny, testowy, czy roboczy(2)), umożliwia wstępną ocenę uzyskanych prognoz — tworzy macierz odwołań (porównanie wartości prognozowanych z prawdziwymi);
- *Text Output* — uruchamia topologię i drukuje wyniki w postaci tekstowej (na ekranie albo do pliku); użytkownik może określić jakie elementy raportu mają się znaleźć w pliku (nagłówek, rekordy, format danych wejściowych i wyników, statystyki);
- *Graphics Output* — narzędzie do graficznej wizualizacji danych, dosyć ubogie i prymitywne, bez możliwości ingerencji w zmianę wyglądu grafiki; jedyną możliwością zapisu takiej wizualizacji jest format bitmapy, lub przenoszenie do innych aplikacji działających w środowisku Windows, poprzez schowek (opcja *Copy/Paste*);
- *The What If?* — proste narzędzie umożliwiające zaobserwowanie wpływu dwóch zmiennych na siebie (np. *co się stanie ze zmienną B, jeśli zmienna A wzrośnie o 5?*), z wykresami czułości i odwołań;
- *The Time Series Plot* — wyświetla równie ubogi graficznie dwuwymiarowy wykres z przebiegami różnych zmiennych: wynik prognozy, dane docelowe–zbiór treningowy, dane docelowe–zbiór walidacyjny, dane docelowe–dane testowe, dane wejściowe (możliwość zapisu wykresu w formacie bitmapy).

Narzędzia do modelowania i prognozowania (Modeling and Forecasting Tools)

W grupie narzędzi do modelowania i prognozowania znajdują się (rys. 2.16):

- *Multi-Layer Perceptron* — wielowarstwowy perceptron, narzędzie do predykcji i klasyfikacji (opisane szczegółowo w rozdz. 2.2.3);
- *Radial Basis Function* — radialna funkcja bazowa, narzędzie do predykcji i klasyfikacji (opisane szczegółowo w rozdz. 2.2.4);
- *Bayesian Network* — sieć Bayesa, pozwalająca na rozwiązywanie zagadnień dotyczących prognozowania i klasyfikacji (patrz rozdz. 2.2.5);
- *Kohonen network* — sieć Kohonena, narzędzie do grupowania (klastrowania) danych;
- *Closest Class Means Classifier* — narzędzie do klasyfikacji;
- *Regression* — regresja, narzędzie do rozwiązywania problemów predykcji;
- *Principal Component Analysis* — analiza głównych składowych, narzędzie, które może być wykorzystane do zredukowania złożoności modelu.

Narzędzia filtrujące (Filter Tools)

W grupie narzędzi filtrujących znajdują się (rys. 2.16):

- *Filter Tool* — narzędzie do wstępnej analizy danych (selekcja danych, ważenie danych, przycinanie, prosta analiza rozkładu, przed i po transformacji);
- *Combiner Tool* — narzędzie pozwalające kombinować wyniki z dwóch lub więcej obiektów w pojedynczy (*single output*), co pozwala na budowanie złożonych modeli;
- *Simulator Tool* — to narzędzie musi być stosowane wraz z narzędziem *What If?* i *Graphics Output*; tworzy tzw. pseudotestowy plik (szczegóły można znaleźć w dokumentacji do programu Neural Connection, SPSS 1997, 1999);
- *Times Series Window* — narzędzie do analizy szeregów czasowych.

(2) W zbiorze roboczym można zamieścić nowe, „świeże" dane, dla których chcemy uzyskać prognozy.

Przykłady rozwiązywania zagadnień predykcji i klasyfikacji w układzie przestrzennym za pomocą programu Neural Connection znajdują się w rozdziale 3. Szczegóły dotyczące programu (opis opcji, procedur, przykłady zastosowań, itp.) można znaleźć w dokumentacji (SPSS, 1997, 1999), na stronach internetowych firmy SPSS (**http://www.spss.com**, **http://www.spss.com.pl**) oraz w systemie pomocy do programu (*Help*).

rozdział 3

Prognozowanie zmian jakości wód podziemnych w układzie przestrzennym

Do prognozowania zmian jakości wód podziemnych dorzecza górnej Wisły w układzie przestrzennym wykorzystane zostaną zweryfikowane wskaźniki fizyko-chemiczne uzyskane w I serii opróbowania sieci RMWP dorzecza górnej Wisły (okres mokry, V–IX 1993).

W serii tej opróbowaniem i analizą objęto 167 punktów RMWP, gdyż punkty 11012, 21024, 21047, 21052 i 21060 (wg numeracji punktów w bazie MONBADA), ze względu na niezakończony proces ich adaptacji nie zostały opróbowane (Witczak et al., 1994).

W zweryfikowanej bazie danych (patrz rozdz. 1.2) pozostało szesnaście zmiennych — wskaźników fizyko-chemicznych jakości wód:

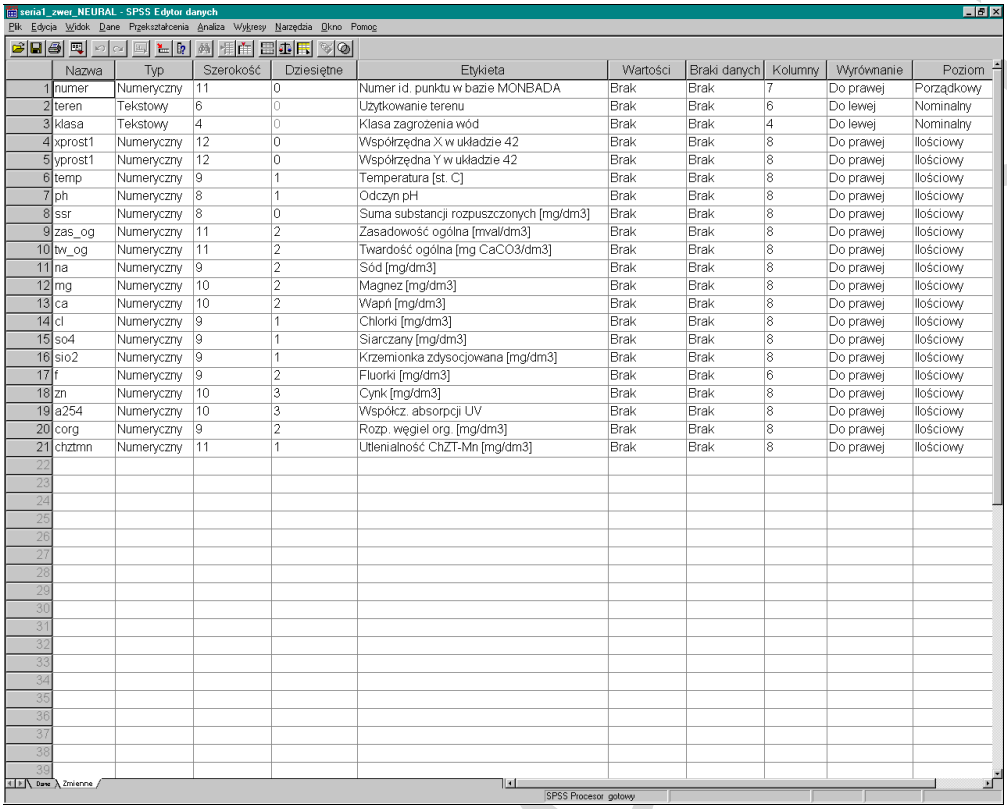
- temperatura [°C];
- odczyn pH;
- suma substancji rozpuszczonych [mg/dm^3];
- zasadowość ogólna [mval/dm^3];
- twardość ogólna [mg $CaCO_3$/dm^3];
- sód [mg/dm^3];
- magnez [mg/dm^3];
- wapń [mg/dm^3];
- chlorki [mg/dm^3];
- siarczany [mg/dm^3];
- krzemionka zdysocjowana [mg/dm^3];
- fluorki [mg/dm^3];
- cynk [mg/dm^3];
- współczynnik absorpcji UV (A 254);
- rozpuszczony węgiel organiczny [mg/dm^3];
- utlenialność ChZT-Mn [mg/dm^3].

Oprócz tych zmiennych w bazie danych umieszczono zmienne umożliwiające identyfikację punktów monitoringowych: numer identyfikacyjny punktu w bazie MONBADA, współrzędne punktu w układzie 42: współrzędna X, współrzędna Y oraz klasę zagrożenia wód (AB, C, D) i sposób użytkowania terenu w otoczeniu danego punktu (R, L, O-P).

Bazę tę zapisano do pliku w formacie SPSS, pod nazwą *seria1_zwer_NEURAL.sav*, plik ten znajduje się na płycie CD-ROM dołączonej do niniejszej pracy. W pliku jest zatem **21 zmiennych** (16 wskaźników fizyko-chemicznych wód i 5 zmiennych typu opisowego — rys. 3.1), opisujących **167 punktów RMWP** (167 wierszy danych — rys. 3.2).

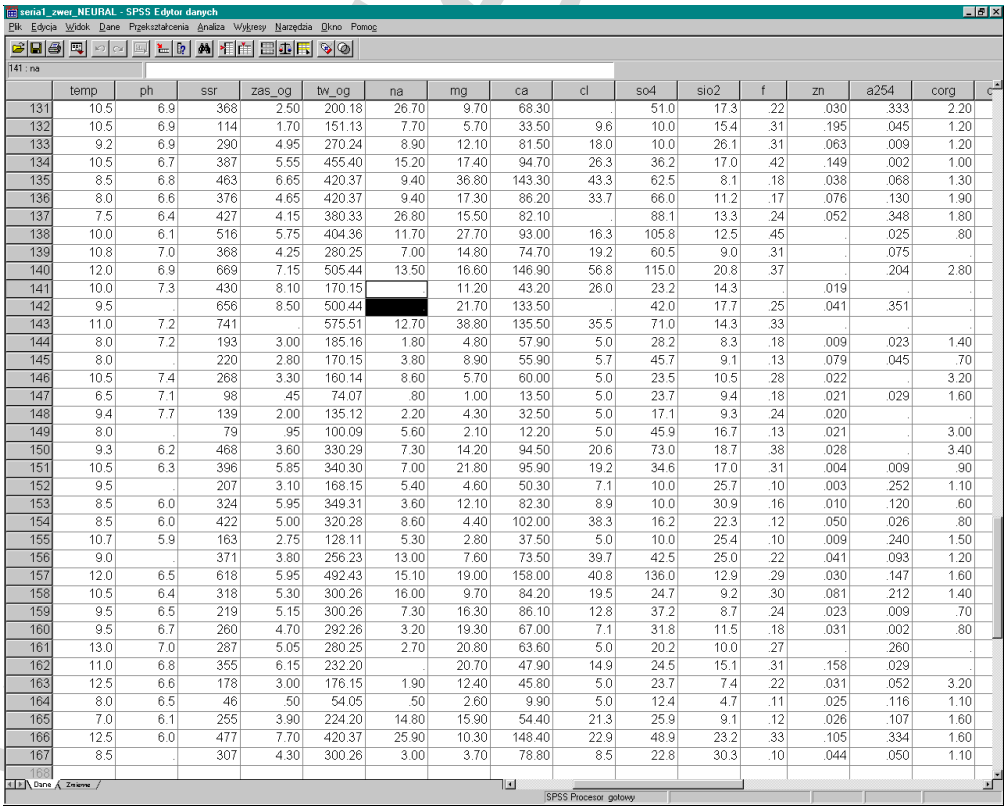
Na tak przygotowanej bazie danych zostaną przeprowadzone próby **predykcji** wskaźników fizyko-chemicznych wód na podstawie współrzędnych punktu monitoringowego oraz **klasyfikacji** punktu monitoringowego (na podstawie wyników oznaczeń wskaźników fizyko-chemicznych) do obszaru o określonym użytkowaniu terenu.

Próby te zostaną przeprowadzone dla trzech wariantów zweryfikowanych danych.



	Nazwa	Typ	Szerokość	Dziesiętne	Etykieta	Wartości	Braki danych	Kolumny	Wyrównanie	Poziom
1	numer	Numeryczny	11	0	Numer id. punktu w bazie MONBADA	Brak	Brak	7	Do prawej	Porządkowy
2	teren	Tekstowy	6	0	Użytkowanie terenu	Brak	Brak	6	Do lewej	Nominalny
3	klasa	Tekstowy	4	0	Klasa zagrożenia wód	Brak	Brak	4	Do lewej	Nominalny
4	xprost1	Numeryczny	12	0	Współrzędna X w układzie 42	Brak	Brak	8	Do prawej	Ilościowy
5	yprost1	Numeryczny	12	0	Współrzędna Y w układzie 42	Brak	Brak	8	Do prawej	Ilościowy
6	temp	Numeryczny	9	1	Temperatura [st. C]	Brak	Brak	8	Do prawej	Ilościowy
7	ph	Numeryczny	8	1	Odczyn pH	Brak	Brak	8	Do prawej	Ilościowy
8	ssr	Numeryczny	8	0	Suma substancji rozpuszczonych [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
9	zas_og	Numeryczny	11	2	Zasadowość ogólna [mval/dm3]	Brak	Brak	8	Do prawej	Ilościowy
10	tw_og	Numeryczny	11	2	Twardość ogólna [mg CaCO3/dm3]	Brak	Brak	8	Do prawej	Ilościowy
11	na	Numeryczny	9	2	Sód [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
12	mg	Numeryczny	10	2	Magnez [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
13	ca	Numeryczny	10	2	Wapń [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
14	cl	Numeryczny	9	1	Chlorki [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
15	so4	Numeryczny	9	1	Siarczany [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
16	sio2	Numeryczny	9	1	Krzemionka zdysocjowana [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
17	f	Numeryczny	9	2	Fluorki [mg/dm3]	Brak	Brak	6	Do prawej	Ilościowy
18	zn	Numeryczny	10	3	Cynk [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
19	a254	Numeryczny	10	3	Współcz. absorpcji UV	Brak	Brak	8	Do prawej	Ilościowy
20	corg	Numeryczny	9	2	Rozp. węgiel org. [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
21	chztmn	Numeryczny	11	1	Utlenialność ChZT-Mn [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy

Rysunek 3.1. Zweryfikowana baza danych — podgląd zmiennych

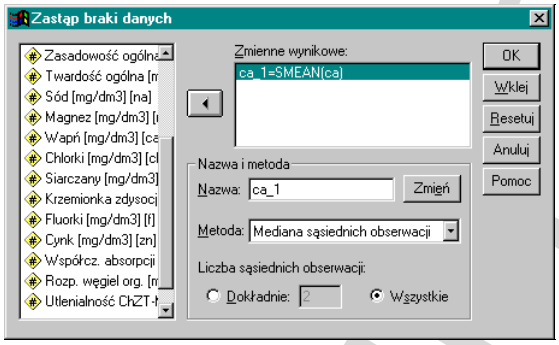


	temp	ph	ssr	zas_og	tw_og	na	mg	ca	cl	so4	sio2	f	zn	a254	corg
131	10.5	6.9	368	2.50	200.18	26.70	9.70	68.30	.	51.0	17.3	.22	.030	.333	2.20
132	10.5	6.9	114	1.70	151.13	7.70	5.70	33.50	9.6	10.0	15.4	.31	.195	.045	1.20
133	9.2	6.9	290	4.95	270.24	8.90	12.10	81.50	18.0	10.0	26.1	.31	.063	.009	1.20
134	10.5	6.7	387	5.55	455.40	15.20	17.40	94.70	26.3	36.2	17.0	.42	.149	.002	1.00
135	8.5	6.8	463	6.65	420.37	9.40	36.80	143.30	43.3	62.5	8.1	.18	.038	.068	1.30
136	8.0	6.6	376	4.65	420.37	9.40	17.30	86.20	33.7	66.0	11.2	.17	.076	.130	1.90
137	7.5	6.4	427	4.15	380.33	26.80	15.50	82.10	.	88.1	13.3	.24	.052	.348	1.80
138	10.0	6.1	516	5.75	404.36	11.70	27.70	93.00	16.3	105.8	12.5	.45	.	.025	.80
139	10.8	7.0	368	4.25	280.25	7.00	14.80	74.70	19.2	60.5	9.0	.31	.	.075	.
140	12.0	6.9	669	7.15	505.44	13.50	16.60	146.90	56.8	115.0	20.8	.37	.	.204	2.80
141	10.0	7.3	430	8.10	170.15	.	11.20	43.20	26.0	23.2	14.3	.	.019	.	.
142	9.5	.	656	8.50	500.44	.	21.70	133.50	.	42.0	17.7	.25	.041	.351	.
143	11.0	7.2	741	.	575.51	12.70	38.80	135.50	35.5	71.0	14.3	.33	.	.	.
144	8.0	7.2	193	3.00	185.16	1.80	4.80	57.90	5.0	28.2	8.3	.18	.009	.023	1.40
145	8.0	.	220	2.80	170.15	3.80	8.90	55.90	5.7	45.7	9.1	.13	.079	.045	.70
146	10.5	7.4	268	3.30	160.14	8.60	5.70	60.00	5.0	23.5	10.5	.28	.022	.	3.20
147	6.5	7.1	98	.45	74.07	.80	1.00	13.50	5.0	23.7	9.4	.18	.021	.029	1.60
148	9.4	7.7	139	2.00	135.12	2.20	4.30	32.50	5.0	17.1	9.3	.24	.020	.	.
149	8.0	.	79	.95	100.09	5.60	2.10	12.20	5.0	45.9	16.7	.13	.021	.	3.00
150	9.3	6.2	468	3.60	330.29	7.30	14.20	94.50	20.6	73.0	18.7	.38	.028	.	3.40
151	10.5	6.3	396	5.85	340.30	7.00	21.80	95.90	19.2	34.6	17.0	.31	.004	.009	.90
152	9.5	.	207	3.10	168.15	5.40	4.60	50.30	7.1	10.0	25.7	.10	.003	.252	1.10
153	8.5	6.0	324	5.95	349.31	3.60	12.10	82.30	8.9	10.0	30.9	.16	.010	.120	.60
154	8.5	6.0	422	5.00	320.28	8.60	4.40	102.00	38.3	16.2	22.3	.12	.050	.026	.80
155	10.7	5.9	163	2.75	128.11	5.30	2.80	37.50	5.0	10.0	25.4	.10	.009	.240	1.50
156	9.0	.	371	3.80	256.23	13.00	7.60	73.50	39.7	42.5	25.0	.22	.041	.093	1.20
157	12.0	6.5	618	5.95	492.43	15.10	19.00	158.00	40.8	136.0	12.9	.29	.030	.147	1.60
158	10.5	6.4	318	5.30	300.26	16.00	9.70	84.20	19.5	24.7	9.2	.30	.081	.212	1.40
159	9.5	6.5	219	5.15	300.26	7.30	16.30	86.10	12.8	37.2	8.7	.24	.023	.009	.70
160	9.5	6.7	260	4.70	292.26	3.20	19.30	67.00	7.1	31.8	11.5	.18	.031	.002	.80
161	13.0	7.0	287	5.05	280.25	2.70	20.80	63.60	5.0	20.2	10.0	.27	.	.260	.
162	11.0	6.8	355	6.15	232.20	.	20.70	47.90	14.9	24.5	15.1	.31	.158	.029	.
163	12.5	6.6	178	3.00	176.15	1.90	12.40	45.80	5.0	23.7	7.4	.22	.031	.052	3.20
164	8.0	6.5	46	.50	54.05	.50	2.60	9.90	5.0	12.4	4.7	.11	.025	.116	1.10
165	7.0	6.1	255	3.90	224.20	14.80	15.90	54.40	21.3	25.9	9.1	.12	.026	.107	1.60
166	12.5	6.0	477	7.70	420.37	25.90	10.30	148.40	22.9	48.9	23.2	.33	.105	.334	1.60
167	8.5	.	307	4.30	300.26	3.00	3.70	78.80	8.5	22.8	30.3	.10	.044	.050	1.10

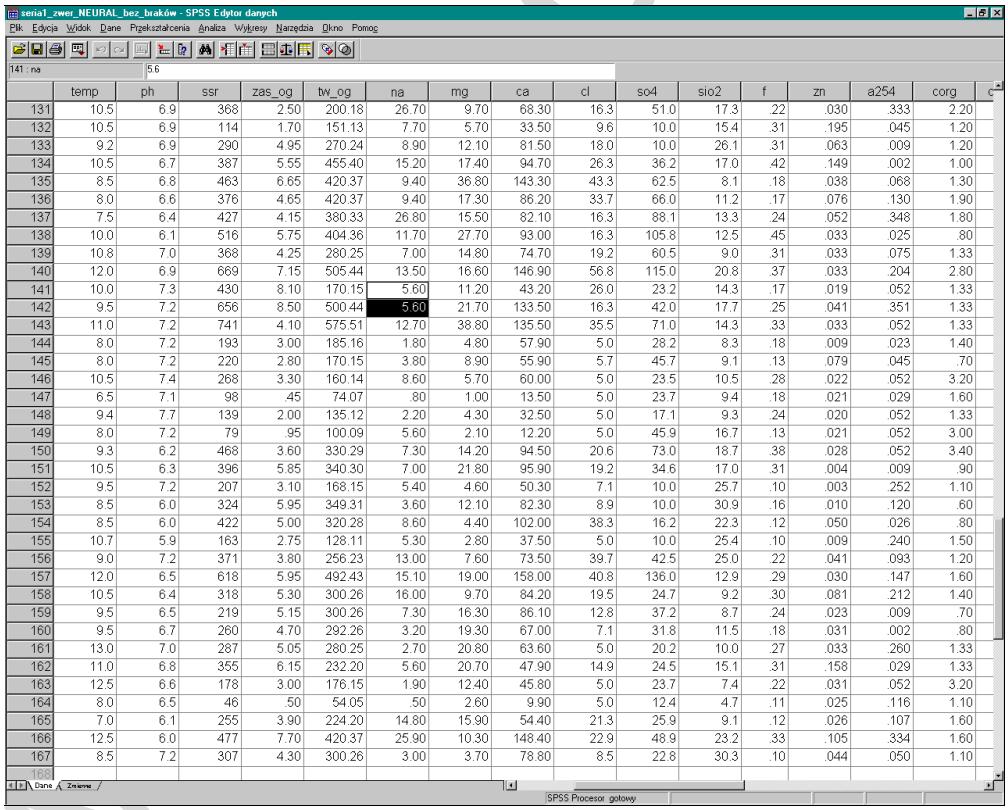
Rysunek 3.2. Zweryfikowana baza danych — podgląd danych. Szarym kolorem podświetlone są przykłady braków danych w zmiennej *sód* [mg/dm^3]

Wariant 1. Plik z danymi reprezentującymi wszystkie klasy zagrożenia wód (AB, C, D) i zagospodarowania terenu (R, L, O-P)

Aby przygotować zbiór danych do wczytania go do programu Neural Connection — dokonano operacji zastąpienia braków danych (w obrębie każdej ze zmiennych reprezentujących wskaźniki fizyko-chemiczne wód) medianą ze wszystkich obserwacji (medianę wybrano ze względu na asymetryczne rozkłady badanych zmiennych). W tym celu wykorzystano opcję programu SPSS **Przekształcenia ▶ Zastąp braki danych** (rys. 3.3). Na rysunku 3.4 widoczny jest efekt działania tej procedury.



Rysunek 3.3. Program SPSS PL for Windows. Opcja zastępowania braków danych medianą ze wszystkich obserwacji dla zmiennej *wapń* [mg/dm^3]



	temp	ph	ssr	zas_og	tw_og	na	mg	ca	cl	so4	sio2	f	zn	a254	corg
131	10.5	6.9	368	2.50	200.18	26.70	9.70	68.30	16.3	51.0	17.3	.22	.030	.333	2.20
132	10.5	6.9	114	1.70	151.13	7.70	5.70	33.50	9.6	10.0	15.4	.31	.195	.045	1.20
133	9.2	6.9	290	4.95	270.24	8.90	12.10	81.50	18.0	10.0	26.1	.31	.063	.009	1.20
134	10.5	6.7	387	5.55	455.40	15.20	17.40	94.70	26.3	36.2	17.0	.42	.149	.002	1.00
135	8.5	6.8	463	6.65	420.37	9.40	36.80	143.30	43.3	62.5	8.1	.18	.038	.068	1.30
136	8.0	6.6	376	4.65	420.37	9.40	17.30	86.20	33.7	66.0	11.2	.17	.076	.130	1.90
137	7.5	6.4	427	4.15	380.33	26.80	15.50	82.10	16.3	88.1	13.3	.24	.052	.348	1.80
138	10.0	6.1	516	5.75	404.36	11.70	27.70	93.00	16.3	105.8	12.5	.45	.033	.025	.80
139	10.8	7.0	368	4.25	280.25	7.00	14.80	74.70	19.2	60.5	9.0	.31	.033	.075	1.33
140	12.0	6.9	669	7.15	505.44	13.50	16.60	146.90	56.8	115.0	20.8	.37	.033	.204	2.80
141	10.0	7.3	430	8.10	170.15	5.60	11.20	43.20	26.0	23.2	14.3	.17	.019	.052	1.33
142	9.5	7.2	656	8.50	500.44	5.60	21.70	133.50	16.3	42.0	17.7	.25	.041	.351	1.33
143	11.0	7.2	741	4.10	575.51	12.70	38.80	135.50	35.5	71.0	14.3	.33	.033	.052	1.33
144	8.0	7.2	193	3.00	185.16	1.80	4.80	57.90	5.0	28.2	8.3	.18	.009	.023	1.40
145	8.0	7.2	220	2.80	170.15	3.80	8.90	55.90	5.7	45.7	9.1	.13	.079	.045	.70
146	10.5	7.4	268	3.30	160.14	8.60	5.70	60.00	5.0	23.5	10.5	.28	.022	.052	3.20
147	6.5	7.1	98	.45	74.07	.80	1.00	13.50	5.0	23.7	9.4	.18	.021	.029	1.60
148	9.4	7.7	139	2.00	135.12	2.20	4.30	32.50	5.0	17.1	9.3	.24	.020	.052	1.33
149	8.0	7.2	79	.95	100.09	5.60	2.10	12.20	5.0	45.9	16.7	.13	.021	.052	3.00
150	9.3	6.2	468	3.60	330.29	7.30	14.20	94.50	20.6	73.0	18.7	.38	.028	.052	3.40
151	10.5	6.3	396	5.85	340.30	7.00	21.80	95.90	19.2	34.6	17.0	.31	.004	.009	.90
152	9.5	7.2	207	3.10	168.15	5.40	4.60	50.30	7.1	10.0	25.7	.10	.003	.252	1.10
153	8.5	6.0	324	5.95	349.31	3.60	12.10	82.30	8.9	10.0	30.9	.16	.010	.120	.60
154	8.5	6.0	422	5.00	320.28	8.60	4.40	102.00	38.3	16.2	22.3	.12	.050	.026	.80
155	10.7	5.9	163	2.75	128.11	5.30	2.80	37.50	5.0	10.0	25.4	.10	.009	.240	1.50
156	9.0	7.2	371	3.80	256.23	13.00	7.60	73.50	39.7	42.5	25.0	.22	.041	.093	1.20
157	12.0	6.5	618	5.95	492.43	15.10	19.00	158.00	40.8	136.0	12.9	.29	.030	.147	1.60
158	10.5	6.4	318	5.30	300.26	16.00	9.70	84.20	19.5	24.7	9.2	.30	.081	.212	1.40
159	9.5	6.5	219	5.15	300.26	7.30	16.30	86.10	12.8	37.2	8.7	.24	.023	.009	.70
160	9.5	6.7	260	4.70	292.26	3.20	19.30	67.00	7.1	31.8	11.5	.18	.031	.002	.80
161	13.0	7.0	287	5.05	280.25	2.70	20.80	63.60	5.0	20.2	10.0	.27	.033	.260	1.33
162	11.0	6.8	355	6.15	232.20	5.60	20.70	47.90	14.9	24.5	15.1	.31	.158	.029	1.33
163	12.5	6.6	178	3.00	176.15	1.90	12.40	45.80	5.0	23.7	7.4	.22	.031	.052	3.20
164	8.0	6.5	46	.50	54.05	.50	2.60	9.90	5.0	12.4	4.7	.11	.025	.116	1.10
165	7.0	6.1	255	3.90	224.20	14.80	15.90	54.40	21.3	25.9	9.1	.12	.026	.107	1.60
166	12.5	6.0	477	7.70	420.37	25.90	10.30	148.40	22.9	48.9	23.2	.33	.105	.334	1.60
167	8.5	7.2	307	4.30	300.26	3.00	3.70	78.80	8.5	22.8	30.3	.10	.044	.050	1.10

Rysunek 3.4. Zweryfikowana baza danych — podgląd danych po zastąpieniu braków danych medianą ze wszystkich obserwacji. Szarym kolorem podświetlone są przykłady zastąpionych braków danych w zmiennej *sód* [mg/dm^3]

W pliku dla wariantu 1. jest zatem 21 zmiennych (kolumn) i 167 obserwacji (wierszy danych, punktów RMWP).

Wariant 2. Plik z danymi reprezentującymi klasę zagrożenia wód AB

W wariancie 1, w pliku występowały punkty o różnej klasie zagrożenia wód (AB, C, D). Aby sprawdzić, jak zmieni się jakość prognoz po ograniczeniu zbioru danych wejściowych do punktów o klasie zagrożenia AB, w wariancie drugim wyłączone zostaną z analizy punkty RMWP o klasie zagrożenia C i D.

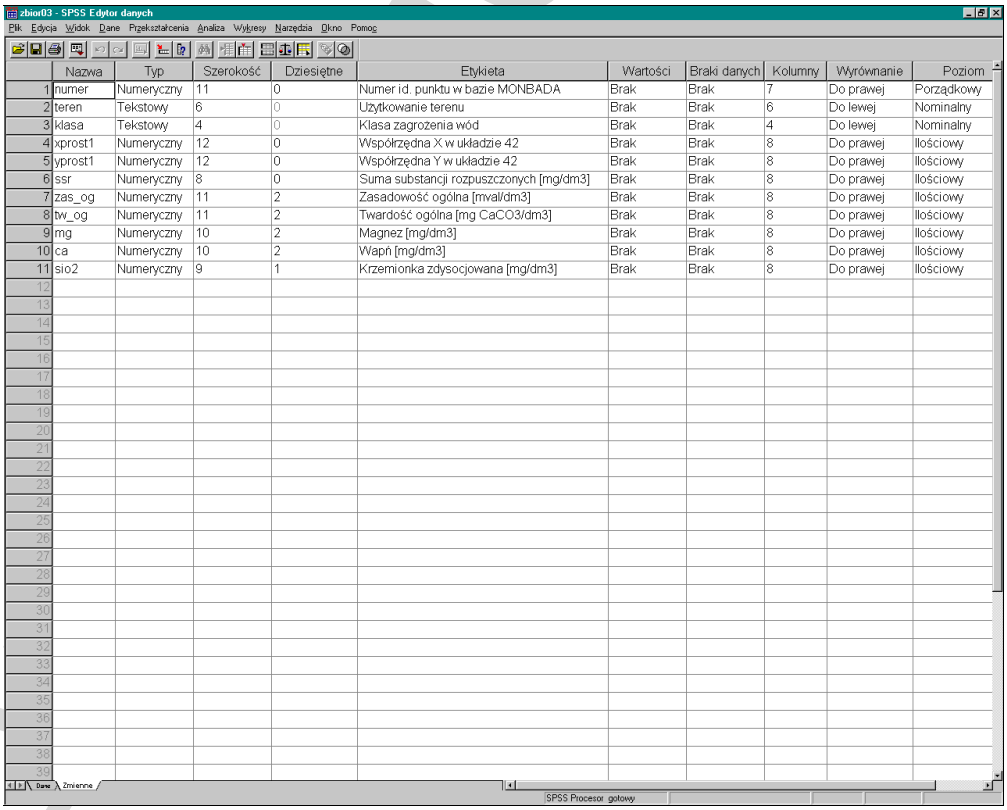
Plik ten będzie miał taką samą konfigurację jak plik w wariancie 1., będzie się składał z 21 zmiennych (16 wskaźników fizyko-chemicznych i 5 zmiennych opisowych), ale do 151 zmniejszy się liczba obserwacji (151 punktów RMWP).

Wariant 3. Plik z danymi reprezentującymi klasę zagrożenia wód AB z ograniczoną liczbą zmiennych (wskaźników fizyko-chemicznych)

W celu sprawdzenia, czy na jakość uzyskiwanych prognoz nie ma wpływu operacja zastępowania znacznej liczby braków danych medianą, w wariancie trzecim będzie testowany plik zawierający punkty o klasie zagrożenia AB, ale ograniczony do 11 zmiennych (6 wskaźników fizyko-chemicznych i 5 zmiennych opisowych).

Spośród zweryfikowanych wskaźników fizyko-chemicznych do pliku wybrano te, w których wystąpiła najmniejsza ($\leqslant 5$) liczba braków danych (tab. 1.13). Są to:

- suma substancji rozpuszczonych [mg/dm^3];
- zasadowość ogólna [mval/dm^3];
- twardość ogólna [mg $CaCO_3$/dm^3];
- magnez [mg/dm^3];
- wapń [mg/dm^3];
- krzemionka zdysocjowana [mg/dm^3] (rys. 3.5).



	Nazwa	Typ	Szerokość	Dziesiętne	Etykieta	Wartości	Braki danych	Kolumny	Wyrównanie	Poziom
1	numer	Numeryczny	11	0	Numer id. punktu w bazie MONBADA	Brak	Brak	7	Do prawej	Porządkowy
2	teren	Tekstowy	6	0	Użytkowanie terenu	Brak	Brak	6	Do lewej	Nominalny
3	klasa	Tekstowy	4	0	Klasa zagrożenia wód	Brak	Brak	4	Do lewej	Nominalny
4	xprost1	Numeryczny	12	0	Współrzędna X w układzie 42	Brak	Brak	8	Do prawej	Ilościowy
5	yprost1	Numeryczny	12	0	Współrzędna Y w układzie 42	Brak	Brak	8	Do prawej	Ilościowy
6	ssr	Numeryczny	8	0	Suma substancji rozpuszczonych [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
7	zas_og	Numeryczny	11	2	Zasadowość ogólna [mval/dm3]	Brak	Brak	8	Do prawej	Ilościowy
8	tw_og	Numeryczny	11	2	Twardość ogólna [mg CaCO3/dm3]	Brak	Brak	8	Do prawej	Ilościowy
9	mg	Numeryczny	10	2	Magnez [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
10	ca	Numeryczny	10	2	Wapń [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy
11	sio2	Numeryczny	9	1	Krzemionka zdysocjowana [mg/dm3]	Brak	Brak	8	Do prawej	Ilościowy

Rysunek 3.5. Zweryfikowana baza danych — plik z ograniczoną liczbą zmiennych

Następnie z pliku wyłączono (usunięto) wszystkie obserwacje z brakami danych — w zbiorze pozostało więc 143 obserwacje (143 punkty RMWP).

W tabeli 3.1 zestawiono konfiguracje plików dla poszczególnych wariantów, wykorzystywanych w dalszej analizie.

Tabela 3.1. Warianty plików do prognozowania zmian jakości wód w układzie przestrzennym w dorzeczu górnej Wisły

Wariant	Liczba zmiennych w pliku	Klasa zagrożenia wód	Liczba punktów RMWP
1.	16 wskaźników fizyko-chemicznych 5 zmiennych typu opisowego	AB, C, D	167
	21 zmiennych		
2.	16 wskaźników fizyko-chemicznych 5 zmiennych typu opisowego	AB	151
	21 zmiennych		
3.	6 wskaźników fizyko-chemicznych 5 zmiennych typu opisowego	AB	143
	11 zmiennych		

3.1. Prognozowanie wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego

W celu stwierdzenia czy na podstawie wyników oznaczeń wskaźników jakości wód w kilku punktach sieci monitoringowej można uzyskać dane dotyczące jakości wód podziemnych we wskazanym punkcie RMWP o znanych współrzędnych, należy zbudować model sieci neuronowej, w której zmiennymi wejściowymi będą współrzędne punktu monitoringowego, a zmiennymi docelowymi — oznaczenia wskaźników fizyko-chemicznych wód podziemnych.

3.1.1. Prognozy dla punktów RMWP reprezentujących wszystkie klasy zagrożenia wód (wariant 1.)

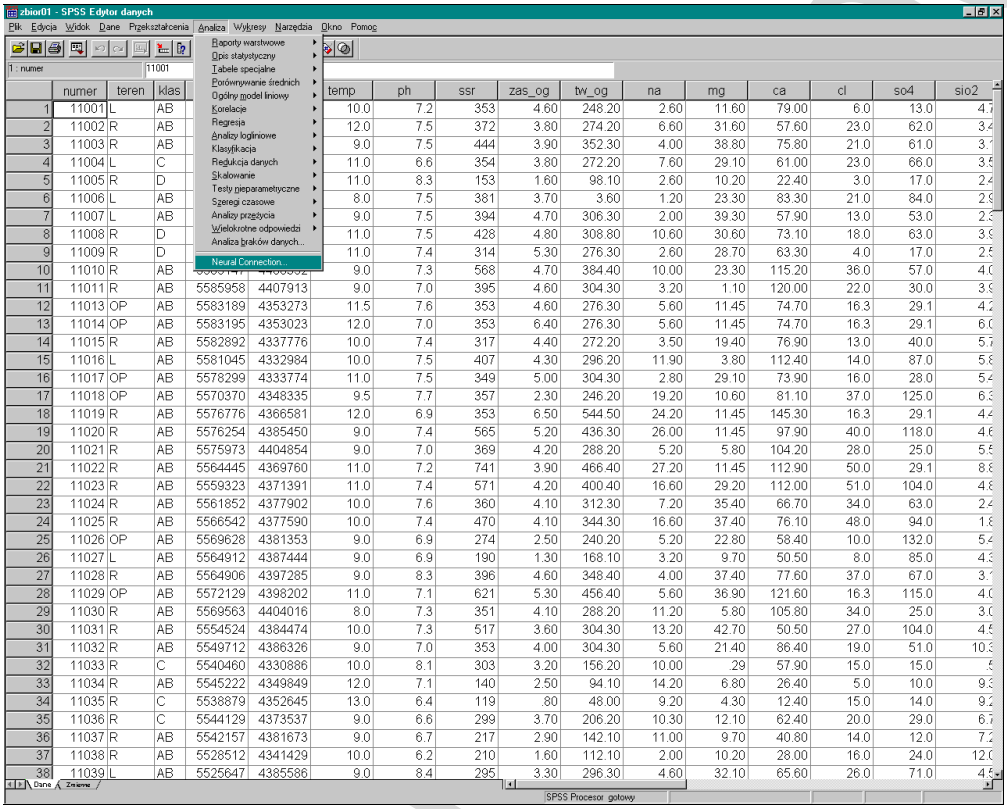
Przygotowany zgodnie z wariantem pierwszym (tab. 3.1) plik danych *zbior01.sav*(1) wczytano wprost do programu Neural Connection, uruchamiając z programu SPSS opcję **Analiza ▶ Neural Connection** (rys. 3.6).

Następnie dokonano konfiguracji zmiennych (rys. 3.7), w taki sposób, że zmienne: numer identyfikacyjny punktu w bazie MONBADA, sposób użytkowania terenu i klasa zagrożenia wód zdefiniowano jako zmienne typu opisowego (R), a współrzędne punktu monitoringowego w układzie 42 jako zmienne typu wejściowego (I), na podstawie których będzie prognozowane 16 zmiennych docelowych (T) — wskaźników fizyko-chemicznych wód (rys. 3.8).

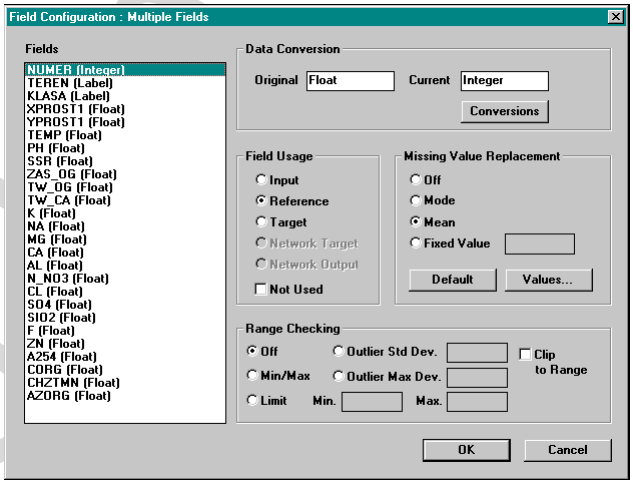
Następnie przystąpiono do budowy modelu sieci. W celu wyboru optymalnej (dającej najlepsze rezultaty prognoz) struktury sieci testowano różne modele z grupy sieci nadzorowanych (*supervised*).

W programie Neural Connection z grupy tej dostępne są sieci: wielowarstwowy perceptron MLP, radialna funkcja bazowa RBF i sieć Bayesa, i one zostały wykorzystane do predykcji wskaźników fizyko-chemicznych wód dla punktu RMWP o znanych współrzędnych.

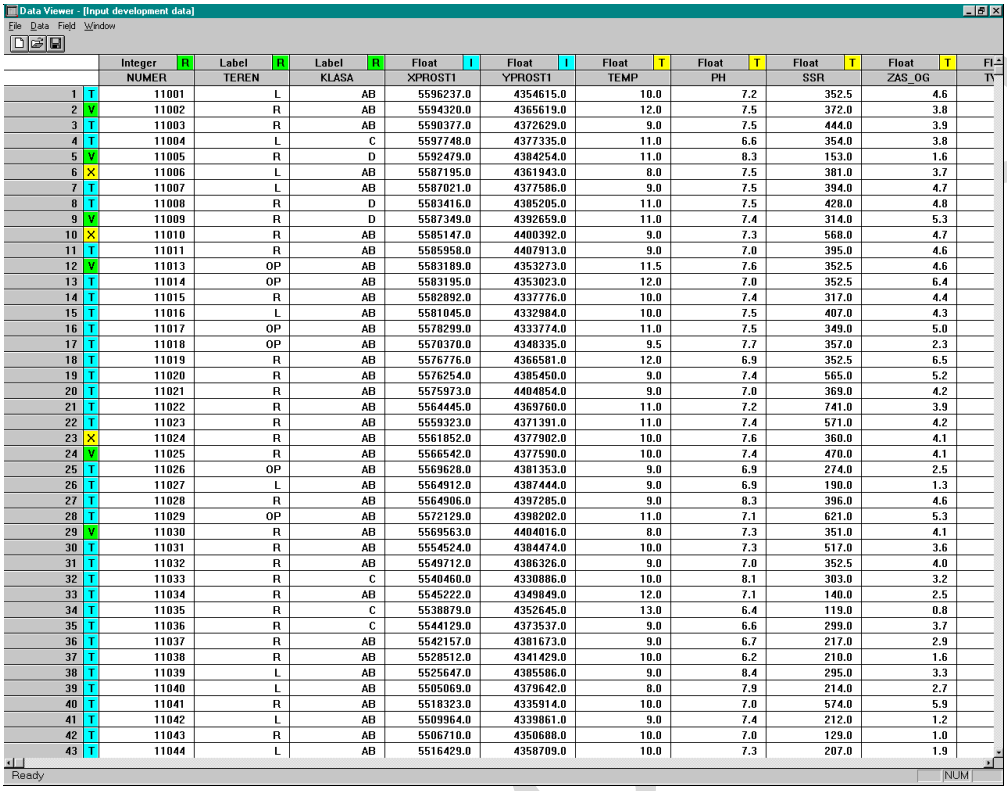
(1) Ten plik, wszystkie pliki z danymi, pliki wynikowe oraz pliki z modelami omawianych sieci neuronowych znajdują się na płycie CD-ROM dołączonej do niniejszej pracy.



Rysunek 3.6. Uruchomienie programu Neural Connection z poziomu programu SPSS

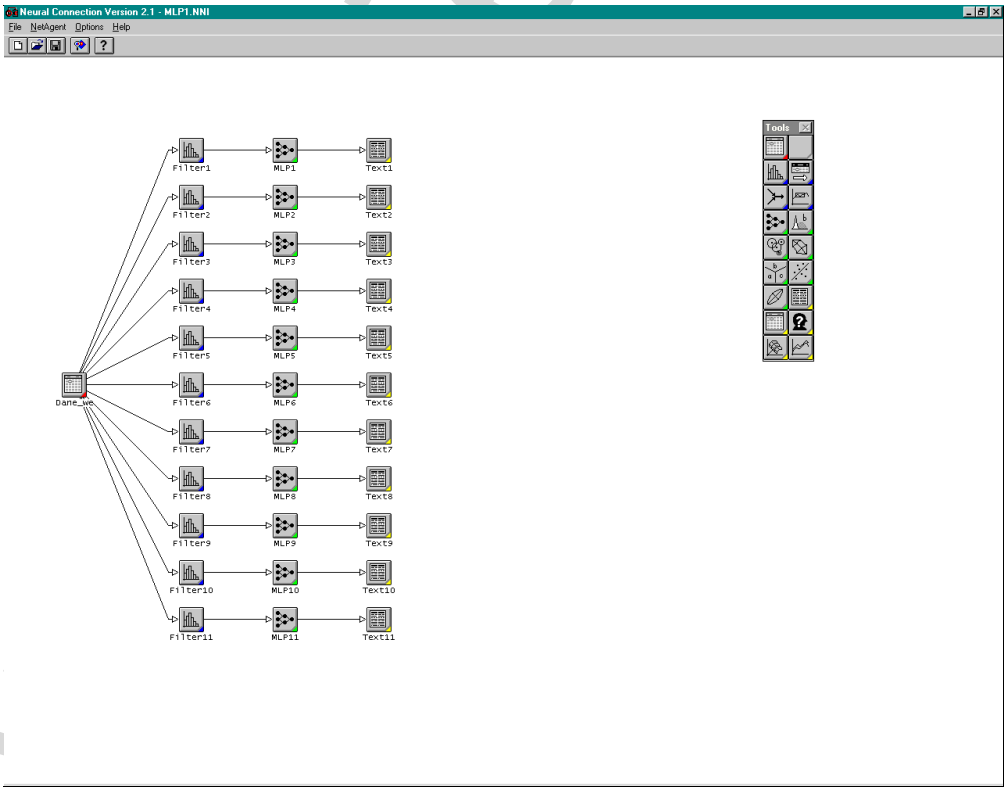


Rysunek 3.7. Ekran konfiguracji zmiennych



		Integer R	Label R	Label R	Float I	Float I	Float T	Float T	Float T	Float T	Fl
		NUMER	TEREN	KLASA	XPROST1	YPROST1	TEMP	PH	SSR	ZAS_OG	TY
1	T	11001	L	AB	5596237.0	4354615.0	10.0	7.2	352.5	4.6	
2	V	11002	R	AB	5594320.0	4365619.0	12.0	7.5	372.0	3.8	
3	T	11003	R	AB	5590377.0	4372629.0	9.0	7.5	444.0	3.9	
4	T	11004	L	C	5597748.0	4377335.0	11.0	6.6	354.0	3.8	
5	V	11005	R	D	5592479.0	4384254.0	11.0	8.3	153.0	1.6	
6	X	11006	L	AB	5587195.0	4361943.0	8.0	7.5	381.0	3.7	
7	T	11007	L	AB	5587021.0	4377586.0	9.0	7.5	394.0	4.7	
8	T	11008	R	D	5583416.0	4385205.0	11.0	7.5	428.0	4.8	
9	V	11009	R	D	5587349.0	4392659.0	11.0	7.4	314.0	5.3	
10	X	11010	R	AB	5585147.0	4400392.0	9.0	7.3	568.0	4.7	
11	T	11011	R	AB	5585958.0	4407913.0	9.0	7.0	395.0	4.6	
12	V	11013	OP	AB	5583189.0	4353273.0	11.5	7.6	352.5	4.6	
13	T	11014	OP	AB	5583195.0	4353023.0	12.0	7.0	352.5	6.4	
14	T	11015	R	AB	5582892.0	4337776.0	10.0	7.4	317.0	4.4	
15	T	11016	L	AB	5581045.0	4332984.0	10.0	7.5	407.0	4.3	
16	T	11017	OP	AB	5578299.0	4333774.0	11.0	7.5	349.0	5.0	
17	T	11018	OP	AB	5570370.0	4348335.0	9.5	7.7	357.0	2.3	
18	T	11019	R	AB	5576776.0	4366581.0	12.0	6.9	352.5	6.5	
19	T	11020	R	AB	5576254.0	4385450.0	9.0	7.4	565.0	5.2	
20	T	11021	R	AB	5575973.0	4404854.0	9.0	7.0	369.0	4.2	
21	T	11022	R	AB	5564445.0	4369760.0	11.0	7.2	741.0	3.9	
22	T	11023	R	AB	5559323.0	4371391.0	11.0	7.4	571.0	4.2	
23	X	11024	R	AB	5561852.0	4377902.0	10.0	7.6	360.0	4.1	
24	V	11025	R	AB	5566542.0	4377590.0	10.0	7.4	470.0	4.1	
25	T	11026	OP	AB	5569628.0	4381353.0	9.0	6.9	274.0	2.5	
26	T	11027	L	AB	5564912.0	4387444.0	9.0	6.9	190.0	1.3	
27	T	11028	R	AB	5564906.0	4397285.0	9.0	8.3	396.0	4.6	
28	T	11029	OP	AB	5572129.0	4398202.0	11.0	7.1	621.0	5.3	
29	V	11030	R	AB	5569563.0	4404016.0	8.0	7.3	351.0	4.1	
30	T	11031	R	AB	5554524.0	4384474.0	10.0	7.3	517.0	3.6	
31	T	11032	R	AB	5549712.0	4386326.0	9.0	7.0	352.5	4.0	
32	T	11033	R	C	5540460.0	4330886.0	10.0	8.1	303.0	3.2	
33	T	11034	R	AB	5545222.0	4349849.0	12.0	7.1	140.0	2.5	
34	T	11035	R	C	5538879.0	4352645.0	13.0	6.4	119.0	0.8	
35	T	11036	R	C	5544129.0	4373537.0	9.0	6.6	299.0	3.7	
36	T	11037	R	AB	5542157.0	4381673.0	9.0	6.7	217.0	2.9	
37	T	11038	R	AB	5528512.0	4341429.0	10.0	6.2	210.0	1.6	
38	T	11039	L	AB	5525647.0	4385586.0	9.0	8.4	295.0	3.3	
39	T	11040	L	AB	5505069.0	4379642.0	8.0	7.9	214.0	2.7	
40	T	11041	R	AB	5518323.0	4335914.0	10.0	7.0	574.0	5.9	
41	T	11042	L	AB	5509964.0	4339861.0	9.0	7.4	212.0	1.2	
42	T	11043	R	AB	5506710.0	4350688.0	10.0	7.0	129.0	1.0	
43	T	11044	L	AB	5516429.0	4358709.0	10.0	7.3	207.0	1.9	

Rysunek 3.8. Ekran podglądu danych wejściowych. Objaśnienia: R — zmienne typu opisowego; I — zmienne typu wejściowego, T — zmienne docelowe

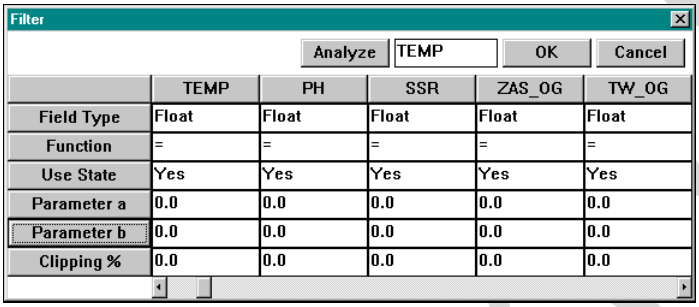


Rysunek 3.9. Schemat sieci typu MLP do prognozowania wskaźników jakości wód podziemnych na podstawie współrzędnych punktu monitoringowego

3.1.1.1. Sieć typu MLP

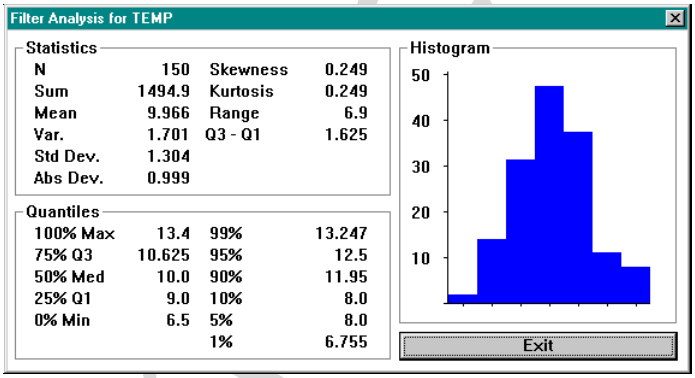
Schemat zbudowanej sieci neuronowej przedstawiony jest na rysunku 3.9.

Narzędzie do filtracji (rys. 3.10) pozwala na ograniczanie liczebności zbioru wejściowego poprzez „odfitrowanie danych" (np. „przycięcie" 5% obserwacji z góry i dołu obserwowanego zakresu zmiennej) lub wyłączenie danej zmiennej z analizy, umożliwia też dokonywanie przekształceń danych (np. operacje logarytmowania).



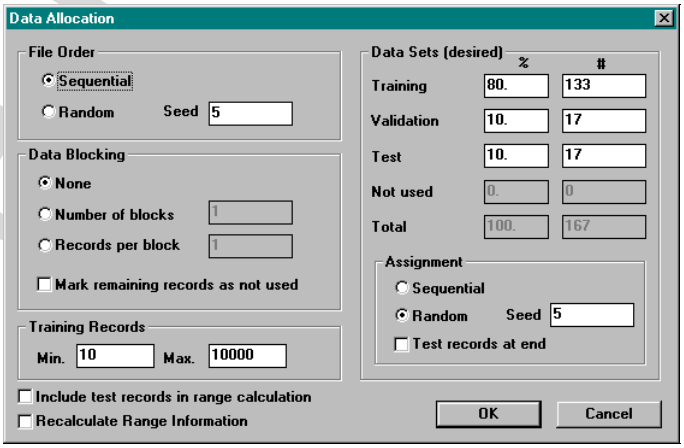
Rysunek 3.10. Narzędzie do filtracji danych

Narzędzie to pozwala również na analizę rozkładu zmiennych (rys. 3.11)



Rysunek 3.11. Narzędzie do filtracji danych — analiza zmiennej wejściowej

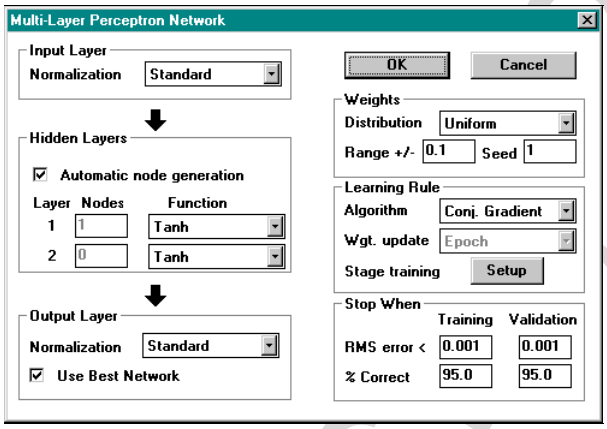
Zbiór danych wejściowych został podzielony na podzbiory (rys. 3.12).



Rysunek 3.12. Ekran podziału danych wejściowych na podzbiory

Podziału dokonano w taki sposób, że w podzbiorze treningowym znajduje się 80% wszystkich obserwacji (133 obserwacje), a w podzbiorach walidacyjnym i testowym po 10% (17 obserwacji).

Przy dwóch zmiennych typu wejściowego ($M = 2$) i szesnastu zmiennych docelowych ($N = 16$) podzbiór treningowy powinien zawierać co najmniej $10(M + N) = 10(2 + 16) = 180$ obserwacji (Tadeusiewicz, 1993; SPSS, 1997). Należy zatem oczekiwać że uzyskane w tym przypadku prognozy mogą nie być zadowalające, z uwagi na mniejszą liczbę obserwacji w zbiorze treningowym (133 obserwacje).

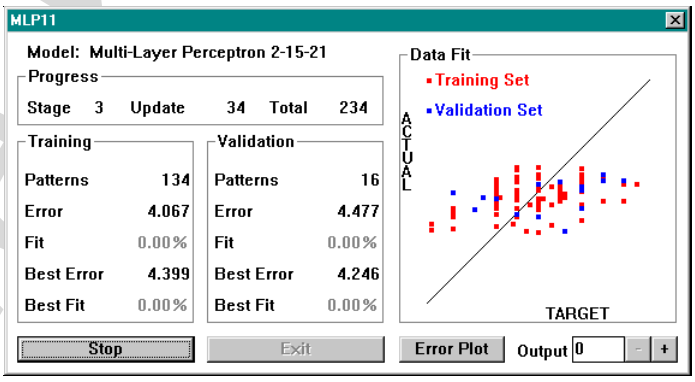


Rysunek 3.13. Okno konfiguracji struktury MLP

Pierwszą próbę „uczenia" sieci przeprowadzono przy domyślnych ustawieniach opcji konfiguracji modułu MLP (rys. 3.13):

- warstwy neuronów wejściowa i wyjściowa są normalizowane, program automatycznie generuje liczbę neuronów w jednej warstwie ukrytej;
- standardowo jako funkcja aktywacji neuronu ustawiony jest tangens hiperboliczny (*tanh*)(2) i jednostajny rozkład wag neuronów (*Uniform*)(3);
- opcja *Use Best Network* umożliwia automatyczny wybór sieci o najlepszych parametrach.

W module MLP są dostępne dwa algorytmy uczące: gradientu sprzężonego (*conj. gradient*) i najszybszego spadku (*steepest descent*). Domyślnie ustawiony jest algorytm gradientu sprzężonego. Szczegóły dotyczące poszczególnych opcji i elementów konfiguracji sieci można znaleźć w dokumentacji do programu Neural Connection (SPSS, 1997, 1999).



Rysunek 3.14. Okno „uczenia się" struktury MLP

(2) Możliwe do wyboru są jeszcze funkcje sigmoid i liniowa.
(3) Opcjonalnie można wybrać rozkład normalny (Gaussa).

Proces „uczenia się" sieci (rys. 3.14) jest zatrzymywany jeśli błąd średniokwadratowy RMS (*Relative Mean Square*) osiągnie wartość mniejszą od 0.001 dla zbiorów treningowego i walidacyjnego, lub inaczej, sieć uzyska 95% poprawnych prognoz w obu zbiorach. Błąd ten liczony jest ze wzoru:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{N}(z_i - y_i)^2}{N}} \tag{3.1}$$

gdzie: z_i — oczekiwany sygnał wyjściowy dla neuronu warstwy wyjściowej; y_i — sygnał wyjściowy pochodzący od neuronu warstwy wyjściowej; N — wielkość warstwy wyjściowej (liczba neuronów).

```
! Input Data Set : Training
!
! ** Data **
!
! Record No. Target Fields                    Output Fields
!            SSR     ZAS_OG    TW_OG     ...  SSR       ZAS_OG    TW_OG     ...
!
      1      372.0   3.8       274.20001      382.8992  3.87707   273.3876 ...
     ...

  Output Error Measures
  =====================
     Output:     RMS Error:      Mean Absolute:  Mean Absolute %:
     -------     ----------      --------------  ---------------
        1        140.423242        108.352583        32.279816 %
        2          1.695153          1.328746        34.157211 %
        3        106.422280         83.914995        32.155673 %
        4         10.101161          8.068774        58.220449 %
        5         32.827660         25.679503        34.338746 %
        6          6.361708          5.076825        41.620086 %
       ...
```

Rysunek 3.15. Fragment zbioru wynikowego *mlp01.nno*

Wyniki „uczenia sieci" dla zbioru treningowego zostały zapisane do zbioru tekstowego *mlp01.nno* (rys. 3.15). W zbiorze tym znajdują się oryginalne zmienne prognozowane, zmienne uzyskane jako odpowiedź sieci oraz parametry określające jakość prognoz dla każdej ze zmiennych — błąd średniokwadratowy (*RMS Error*), odchylenie przeciętne MA (*Mean Absolute*):

$$\mathrm{MA} = \frac{\sum_{i=1}^{N}(z_i - y_i)}{N} \tag{3.2}$$

i średni błąd względny w procentach, liczony jako moduł różnicy pomiędzy wartością prawdziwą a wartością prognozy MA% (*Mean Absolute %*):

$$\mathrm{MA\%} = \mathrm{MA} \cdot \frac{N}{\sum_{i=1}^{N} z_i} \cdot 100\% \tag{3.3}$$

oznaczenia jak we wzorze (3.1).

Ten ostatni parametr podawany jest w zestawieniach, w celu porównania prognoz uzyskanych za pomocą sieci o różnej konfiguracji. Wyniki prognoz zawarte w zbiorze *mlp01.nno* przedstawiono w tabeli 3.2.

Tabela 3.2. Błędy względne prognoz MA% (liczone wg wzoru (3.3)) wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu MLP, zbiór treningowy, punkty klasy AB, C, D (16 zmiennych docelowych)

		Wartość błędu prognozy dla wskaźnika jakości wód [%]															
Lp.	Zbiór wynikowy	Temperatura	Odczyn pH	Suma subst. rozp.	Zasadowość ogólna	Twardość ogólna	Sód	Magnez	Wapń	Chlorki	Siarczany	Krzemionka	Fluorki	Cynk	Współcz. absorpcji UV	Rozpuszczony węgiel org.	Utlenialność ChZT-Mn
1.	*mlp01.nno*	9.92	4.79	34.17	36.86	34.75	63.02	58.99	36.19	54.36	60.09	38.28	44.30	61.37	70.55	40.85	37.55
2.	*mlp02.nno*	9.94	4.69	33.62	36.19	34.22	62.69	58.83	35.35	53.61	59.73	37.88	43.96	61.72	69.65	40.71	37.25
3.	*mlp03.nno*	9.69	4.53	28.47	30.87	27.70	55.72	52.57	31.27	48.67	54.05	34.44	42.22	58.81	68.68	38.52	34.47
4.	*mlp04.nno*	9.99	4.59	28.88	30.38	27.09	56.37	54.42	31.02	50.16	52.53	37.24	40.65	64.48	66.80	36.28	36.03
5.	*mlp05.nno*	9.83	4.80	31.51	34.66	32.48	61.75	57.93	33.71	51.76	57.74	45.07	50.24	63.06	72.47	41.18	36.72
6.	*mlp06.nno*	9.84	4.78	28.88	30.82	27.61	60.61	56.15	29.24	51.67	58.18	33.80	39.91	60.57	68.23	37.20	35.08
7.	*mlp07.nno*	9.91	4.85	33.83	36.18	34.58	63.10	58.93	35.45	53.92	59.76	38.61	44.27	61.67	70.12	40.19	37.08
8.	*mlp08.nno*	9.62	4.47	28.34	28.97	27.75	57.75	53.51	30.01	47.32	54.19	30.23	37.71	58.39	65.22	36.89	34.96
9.	*mlp09.nno*	10.16	4.89	33.86	36.52	34.95	62.86	59.14	36.31	56.00	61.53	44.42	48.49	67.72	71.95	41.64	36.30
10.	*mlp10.nno*	10.02	4.86	34.11	37.51	35.22	63.05	59.15	36.41	55.11	60.25	42.28	45.94	63.91	70.36	40.40	36.92
11.	*mlp11.nno*	9.90	4.85	34.12	36.60	34.89	63.54	58.63	35.99	53.71	59.99	39.52	44.32	62.34	70.28	40.99	37.75

Kolejnym krokiem było testowanie zbudowanego modelu poprzez zmianę parametrów sieci, w celu wyboru zestawu parametrów, dla których uzyskane błędy prognoz będą jak najmniejsze. Najpierw dokonano zmiany funkcji aktywacji neuronu, na funkcję sigmoidalną — uzyskane dla zbioru treningowego wyniki zostały zapisane w zbiorze *mlp02.nno*.

Następne modyfikacje dotyczyły algorytmu uczącego i sposobu uaktualniania wag neuronów, próbowano dołożyć drugą warstwę ukrytą (co znacznie wydłużyło sam proces „uczenia się" sieci — z ok. 3 minut, do 8–10 minut, na komputerze typu Pentium II 266 MHz, 64 MB RAM).

Konfiguracje poszczególnych modeli sieci (opcje z rys. 3.13), i uzyskane dla zbioru treningowego wyniki prognoz wskaźników chemicznych wód zestawione są w tabeli 3.2.

Charakterystyka struktur sieci zestawionych w tabeli 3.2:

- *mlp01.nno* — struktura sieci: 2–12–16 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp02.nno* — struktura sieci: 2–12–16; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp03.nno* — struktura sieci: 2–12–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: stopniowe opadanie (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp04.nno* — konfiguracja jak w przypadku *mlp03.nno*, zmieniona została metoda opcja uaktualniania wag — wagi są uaktualniane po każdym kolejnym etapie (*pattern*);
- *mlp05.nno* — struktura sieci: 2–12–16; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp06.nno* — struktura sieci: 2–12–12–16 (dołożona została druga warstwa ukryta z 12 neuronami); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężeony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); zmiana ta spowodowała znaczne wydłużenie czasu „uczenia się" sieci, bez poprawy jakości uzyskanych wyników;
- *mlp07.nno* — struktura sieci: 2–12–12–16; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp08.nno* — struktura sieci: 2–12–12–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp09.nno* — struktura sieci: 2–16–16 (sieć z jedną warstwą ukrytą, zmiana liczby neuronów w tej warstwie); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); w tym przypadku odnotowano znacznie dłuższy czas uczenia się sieci;
- *mlp10.nno* — struktura sieci: 2–32–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp11.nno* — struktura sieci: 2–12–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: normalny (*gaussian*).

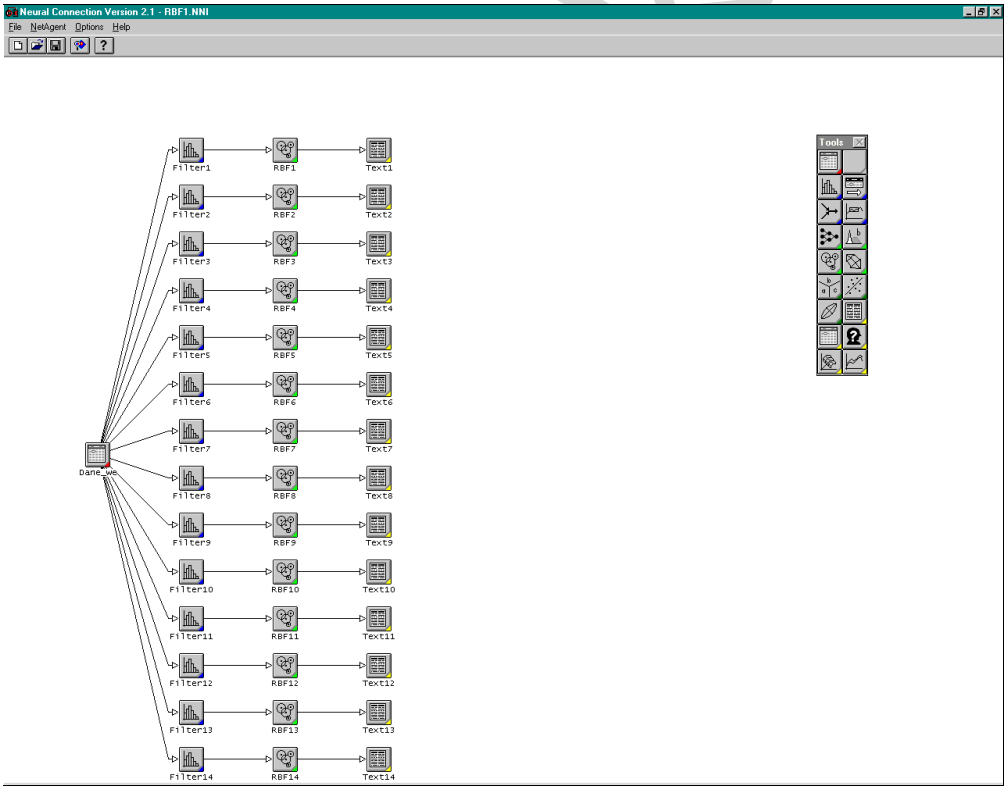
Próby zmiany konfiguracji sieci — dokładanie warstw ukrytych, zmiana liczby neuronów w warstwie ukrytej, modyfikacja rozkładu wag neuronów — nie dały pozytywnych rezultatów (poprawy uzyskanych prognoz).

Błędy względne prognoz MA% kształtowały się na różnym poziomie, od kilku (np. temperatura, odczyn pH) do kilkudziesięciu procent (pozostałe prognozowane wskaźniki fizyko-chemiczne). Nie zaobserwowano związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników (tab. 1.13)

Najlepsze wyniki uzyskano dla struktury 2–12–16 (plik *mlp03.nno*): funkcja aktywacji — tangens hiperboliczny (*tanh*); algorytm uczący — najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów — jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej.

3.1.1.2. Sieć typu RBF

Schemat zbudowanej sieci neuronowej typu RBF przedstawiony jest na rysunku 3.16.

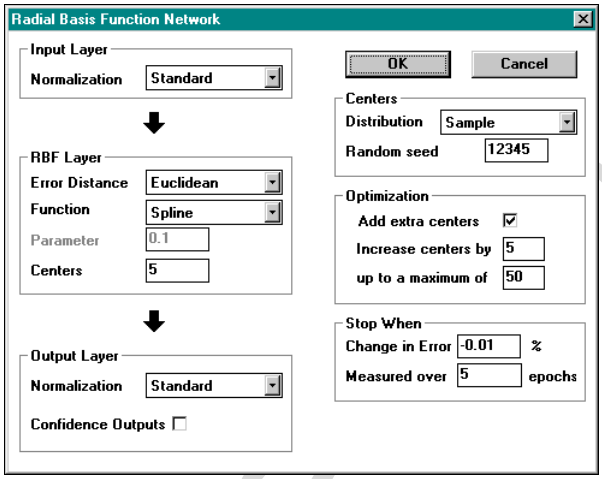


Rysunek 3.16. Schemat sieci RBF do prognozowania wskaźników jakości wód podziemnych na podstawie współrzędnych punktu monitoringowego

Zachowano podział zbioru danych wejściowych jak w przypadku sieci MLP: podzbiór treningowy (80% wszystkich obserwacji — 133 obserwacje), walidacyjny i testowy (po 10% — 17 obserwacji). Pierwszą próbę „uczenia" sieci przeprowadzono przy domyślnych ustawieniach modułu (rys. 3.17):

- standardowa normalizacja warstw wejściowej i wyjściowej;
- liczba centrów: 5;
- miara odległości błędu: Euklidesowa (*Euclidean*); można wybrać też odległość typu *City Block*;

- funkcja nieliniowa: *Spline*, funkcja ta wyraża się wzorem $d^2 \log d$; ponadto dostępne są funkcje: *Gaussa* $(-d^2/(2x\beta^2))$, *Multi-quadratic* $((d^2+\beta^2)^{1/2})$, *Inverse quadratic* $(1/(d^2+\beta^2)^{1/2})$, gdzie: d — odległość od centrum; β — parametr funkcji;
- rozkład centrów: w oparciu o dane (*Sample*) — centra funkcji bazowych znajdują się w wybranych punktach danych; inne dostępne tu opcje to: losowy wybór centrów (*Random*) — centra rozmieszczone są przypadkowo (losowo) w przestrzeni danych, lub tzw. próbny wybór centrów (*Trial*), w którym pierwsze centrum umieszczone jest w dowolnym punkcie danych a pozostałe w skrajnych punktach danych;
- optymalizacja: zwiększanie liczby centrów o 5 aż do 50;
- warunki zakończenia procesu uczenia: zmiana błędu -0.01% w ciągu 5 cykli (epok).



Rysunek 3.17. Okno konfiguracji struktury RBF

Sieć typu RBF uczy się znacznie szybciej niż sieć MLP. Wyniki prognoz przy domyślnych ustawieniach modułu RBF zapisane zostały do zbioru *rbf01.nno*.

Dalsze próby prowadzono przy zmodyfikowanej strukturze sieci RBF, modyfikacje dotyczyły odległości (*error distance*), rodzaju funkcji nieliniowej i jej parametrów. Konfiguracje poszczególnych modeli sieci, i uzyskane wyniki prognoz zestawione są w tabeli 3.3.

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.3:

- *rbf01.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf02.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf03.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf04.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: oparty na danych *Sample*;
- *rbf05.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.3; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf06.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.5; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf07.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf08.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf09.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: próbny (*Trial*);

Tabela 3.3. Błędy względne prognoz MA% (liczone wg wzoru (3.3)) wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór treningowy, punkty klasy AB, C, D (16 zmiennych docelowych)

		Wartość błędu prognozy dla wskaźnika jakości wód [%]															
Lp.	Zbiór wynikowy	Temperatura	Odczyn pH	Suma subst. rozp.	Zasadowość ogólna	Twardość ogólna	Sód	Magnez	Wapń	Chlorki	Siarczany	Krzemionka	Fluorki	Cynk	Współcz. absorpcji UV	Rozpuszczony węgiel org.	Utlenialność ChZT-Mn
1.	*rbf01.nno*	9.65	4.24	29.66	31.80	29.18	55.41	54.64	31.37	50.79	53.05	35.04	40.99	60.71	64.68	35.43	32.19
2.	*rbf02.nno*	9.22	3.83	26.68	26.62	24.28	49.46	48.80	26.41	44.86	49.39	30.83	38.21	58.45	61.70	33.02	30.82
3.	*rbf03.nno*	9.26	4.81	33.40	31.99	30.88	55.09	52.32	35.10	52.99	56.78	45.28	44.80	62.21	66.91	35.33	32.65
4.	*rbf04.nno*	9.33	4.69	32.54	30.91	30.79	53.23	52.17	35.07	52.37	56.09	43.03	43.17	61.84	64.42	34.72	31.79
5.	*rbf05.nno*	9.41	4.64	31.68	30.65	30.19	52.92	51.79	34.81	51.52	55.18	41.71	41.90	61.22	63.39	34.73	31.48
6.	*rbf06.nno*	9.84	4.78	30.98	34.05	31.73	55.54	56.29	35.65	52.02	55.42	42.38	44.35	62.13	69.18	36.37	33.18
7.	*rbf07.nno*	9.86	5.00	33.04	36.36	34.12	62.38	55.94	36.14	53.67	56.04	42.06	47.48	60.85	71.64	40.13	36.96
8.	*rbf08.nno*	8.95	4.40	25.04	26.06	24.27	50.72	50.66	26.91	44.60	53.75	39.58	39.02	56.28	55.48	34.94	30.11
9.	*rbf09.nno*	9.92	4.69	35.87	38.73	36.52	61.16	58.70	39.52	54.74	59.24	43.63	50.11	64.35	68.26	40.11	36.67
10.	*rbf10.nno*	9.76	4.37	29.67	29.54	28.95	53.22	52.69	31.76	50.77	55.82	34.24	40.57	58.67	66.19	34.37	32.45
11.	*rbf11.nno*	10.05	4.68	31.14	34.51	31.87	60.08	55.71	34.40	52.00	57.03	38.99	43.22	61.80	69.26	40.37	36.61
12.	*rbf12.nno*	9.96	4.93	35.97	38.67	36.24	61.44	59.07	39.70	54.99	59.30	44.12	50.93	64.87	67.70	39.72	36.54
13.	*rbf13.nno*	9.84	4.98	33.68	37.28	34.74	62.72	55.48	37.02	53.96	55.93	40.88	47.06	60.74	72.36	40.18	36.89
14.	*rbf14.nno*	8.09	4.51	24.81	26.33	23.89	46.27	48.09	26.90	42.42	49.09	39.45	36.14	55.14	57.61	34.49	29.44
15.	*rbf15.nno*	8.39	4.24	23.61	22.82	22.27	47.28	49.81	24.89	44.29	51.39	35.58	33.90	57.35	56.64	34.50	28.32

- *rbf10.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf11.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf12.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf13.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf14.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf15.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: próbny (*Trial*).

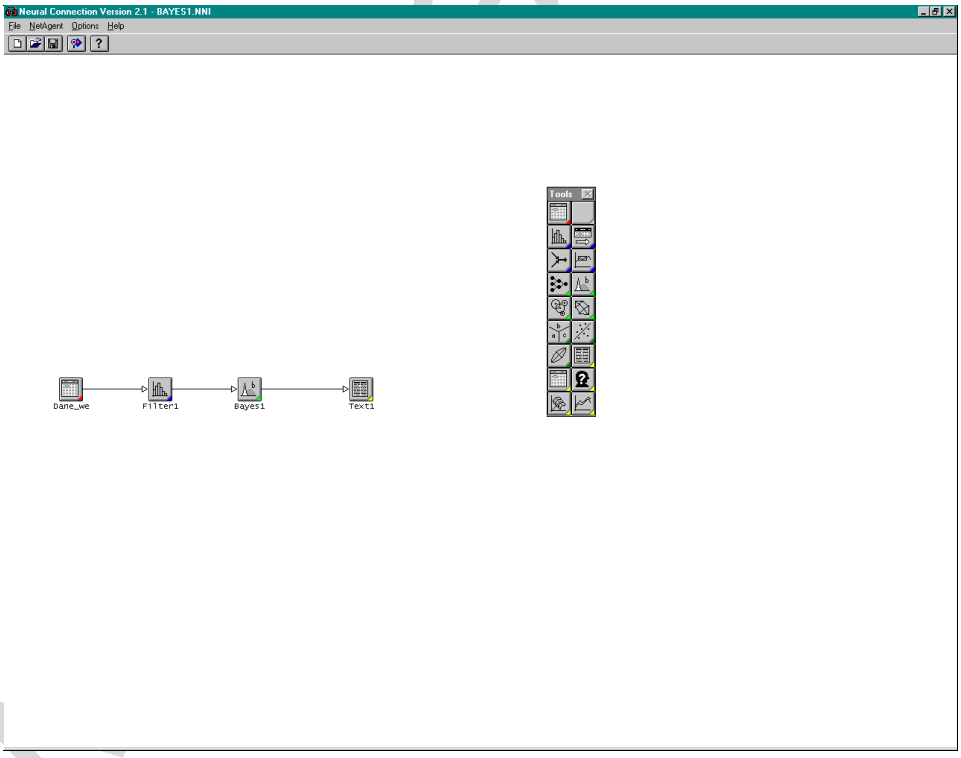
Błędy względne prognoz wskaźników fizyko-chemicznych wód uzyskane za pomocą sieci RBF kształtują się na poziomie od kilku do kilkudziesięciu procent. Sieć typu RBF daje jednak lepsze wyniki prognoz (mniejszy błąd względny prognoz) niż sieć MLP (i w znacznie krótszym czasie).

Podobnie jak w przypadku sieci MLP nie obserwuje się związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników (tab. 1.13).

Najlepsze wyniki prognoz (najmniejsze błędy względne prognoz) uzyskano dla modelu, którego wyniki zapisane są w pliku *rbf15.nno*.

3.1.1.3. Sieć typu Bayesa

Schemat zbudowanej sieci neuronowej typu Bayesa przedstawiony jest na rysunku 3.18.



Rysunek 3.18. Schemat sieci Bayesa do prognozowania wskaźników jakości wód podziemnych na podstawie współrzędnych punktu monitoringowego

Zachowano podział zbioru danych wejściowych jak w przypadku sieci MLP i RBF: podzbiór treningowy (80% wszystkich obserwacji — 133 obserwacje), walidacyjny (10% — 17 ob-

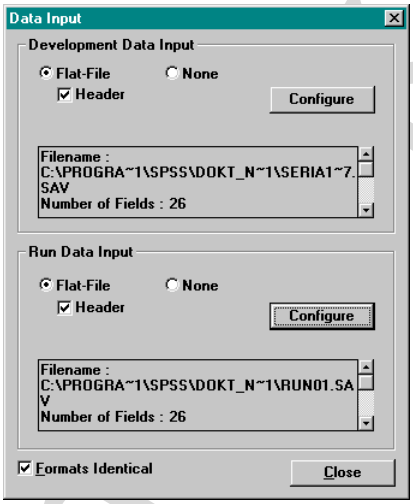
serwacji) i testowy (17 obserwacji). Wybrano opcję wykorzystania przez sieć zbioru walidacyjnego w procesie „uczenia się" sieci.

Nie udało się jednak uzyskać wyników prognoz dla tego typu sieci, gdyż na pewnym etapie treningu program „zawieszał się", niezależnie od zmiany parametrów sieci — błąd treningu cały czas bardzo szybko wzrastał.

3.1.1.4. Wybór najlepszego modelu sieci

Porównując wyniki prognoz uzyskanych za pomocą modeli MLP i RBF (tab. 3.2, tab. 3.3), nalepszym modelem (najmniejsze błędy względne prognoz badanych wskaźników) okazał się model sieci RBF, którego wyniki zapisane są w pliku *rbf15.nno* (odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: *Trial)*.

W celu sprawdzenia zdolności prognozowania „nauczonego" modelu, do struktury wprowadzono dane zewnętrzne, plik roboczy (*run*) *run01.sav* z tymi samymi danymi, na których sieć się uczyła (rys. 3.19).



Rysunek 3.19. Okno do konfiguracji pliku roboczego (*run*)

Wyniki prognoz wskaźników fizyko-chemicznych jakości wód na podstawie współrzędnych punktu monitoringowego dla danych z pliku roboczego zostały zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS *output1.sav* (rys. 3.20, 3.21).

W programie SPSS zostały obliczone błędy względne prognoz B dla każdego z prognozowanych wskaźników:

$$B = \frac{x_{obs} - x_{progn}}{x_{obs}} \cdot 100\% \tag{3.4}$$

gdzie: x_{obs} — wartość obserwowana; x_{progn} — wartość prognozowana.

Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych. Pełny raport z tej części analizy znajduje się na płycie CD-ROM dołączonej do pracy, w pliku *output1.spo*.

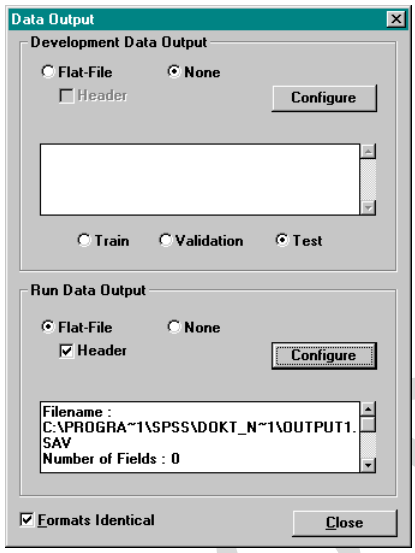
Na rysunku 3.22 przedstawiono histogram rozkładu błędów i wykres rozrzutu wartości obserwowanych i prognozowanych dla cynku.

W tabeli 3.4 zestawione są średnie błędy prognoz B (obliczone wg wzoru (3.4)) dla analizowanych zmiennych — wskaźników fizyko-chemicznych oraz współczynniki korelacji wartości obserwowanych i prognozowanych.

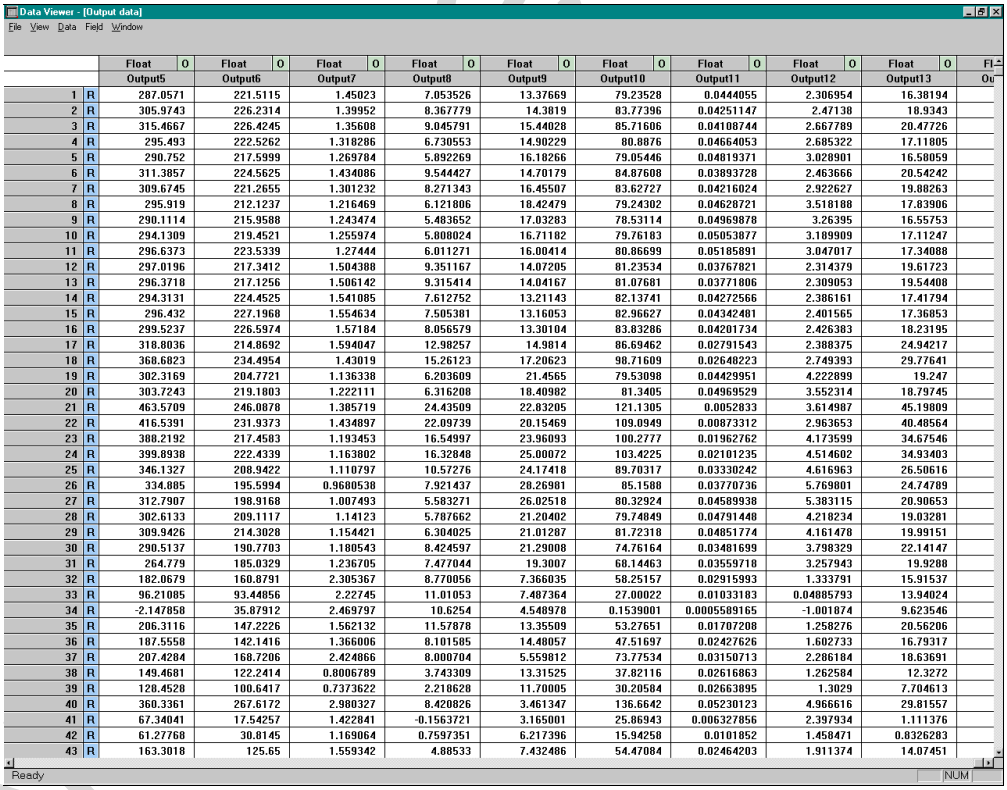
W przypadku cynku, który charakteryzował się najniższą precyzją (tab. 1.13) średni błąd względny prognoz ma niską wartość 3.63%, jednak obserwuje się duży rozrzut tych błędów,

w zakresie od −100 do 280%. Współczynnik korelacji wartości obserwowanych i prognozowanych wynosi zaledwie 0.383.

Średni błąd względny prognoz 16 wskaźników fizyko-chemicznych wód kształtuje się na niskim poziomie, rzędu setnych procenta–kilkunastu procent.



Rysunek 3.20. Okno do konfiguracji zapisu wyników prognoz dla pliku roboczego do pliku zewnętrznego w formacie SPSS

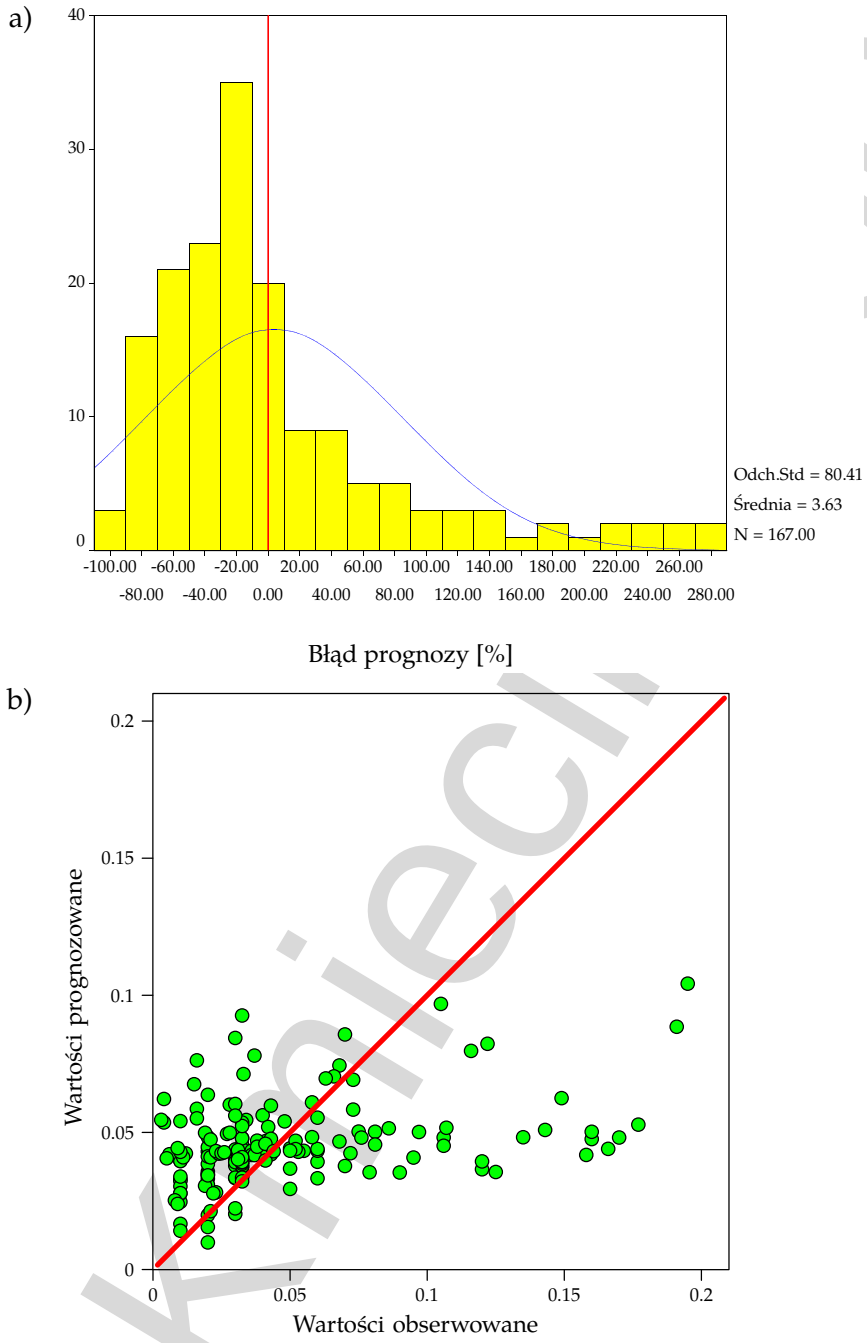

Data Viewer - [Output data]

File View Data Field Window

		Output5	Output6	Output7	Output8	Output9	Output10	Output11	Output12	Output13
1	R	287.0571	221.5115	1.45023	7.053526	13.37669	79.23528	0.0444055	2.306954	16.38194
2	R	305.9743	226.2314	1.39952	8.367779	14.3819	83.77396	0.04251147	2.47138	18.9343
3	R	315.4667	226.4245	1.35608	9.045791	15.44028	85.71606	0.04108744	2.667789	20.47726
4	R	295.493	222.5262	1.318286	6.730553	14.90229	80.8876	0.04664053	2.685322	17.11805
5	R	290.752	217.5999	1.269784	5.892269	16.18266	79.05446	0.04819371	3.028901	16.58059
6	R	311.3857	224.5625	1.434086	9.544427	14.70179	84.87608	0.03893728	2.463666	20.54242
7	R	309.6745	221.2655	1.301232	8.271343	16.45507	83.62727	0.04216024	2.922627	19.88263
8	R	295.919	212.1237	1.216469	6.121806	18.42479	79.24302	0.04628721	3.518188	17.83906
9	R	290.1114	215.9588	1.243474	5.483652	17.03283	78.53114	0.04969878	3.26395	16.55753
10	R	294.1309	219.4521	1.255974	5.808024	16.71182	79.76183	0.05053877	3.189909	17.11247
11	R	296.6373	223.5339	1.27444	6.011271	16.00414	80.86699	0.05185891	3.047017	17.34088
12	R	297.0196	217.3412	1.504388	9.351167	14.07205	81.23534	0.03767821	2.314379	19.61723
13	R	296.3718	217.1256	1.506142	9.315414	14.04167	81.07681	0.03771806	2.309053	19.54408
14	R	294.3131	224.4525	1.541085	7.612752	13.21143	82.13741	0.04272566	2.386161	17.41794
15	R	296.432	227.1968	1.554634	7.505381	13.16053	82.96627	0.04342401	2.401565	17.36853
16	R	299.5237	226.5974	1.57184	8.056579	13.30104	83.83286	0.04201734	2.426383	18.23195
17	R	318.8036	214.8692	1.594047	12.98257	14.9814	86.69462	0.02791543	2.388375	24.94217
18	R	368.6823	234.4954	1.43019	15.26123	17.20623	98.71609	0.02648223	2.749393	29.77641
19	R	302.3169	204.7721	1.136338	6.203609	21.4565	79.53098	0.04429951	4.222899	19.247
20	R	303.7243	219.1803	1.222111	6.316208	18.40982	81.3405	0.04969529	3.552314	18.79745
21	R	463.5709	246.0878	1.385719	24.43509	22.83205	121.1305	0.0052833	3.614987	45.19809
22	R	416.5391	231.9373	1.434897	22.09739	20.15469	109.0949	0.00873312	2.963653	40.48564
23	R	388.2192	217.4583	1.193453	16.54997	23.96093	100.2777	0.01962762	4.173599	34.67546
24	R	399.8938	222.4339	1.163802	16.32848	25.00072	103.4225	0.02101235	4.514602	34.93403
25	R	346.1327	208.9422	1.110797	10.57276	24.17418	89.70317	0.03330242	4.616963	26.50616
26	R	334.885	195.5994	0.9680538	7.921437	28.26981	85.1588	0.03770736	5.769801	24.74789
27	R	312.7907	198.9168	1.007493	5.583271	26.02518	80.32924	0.04589938	5.383115	20.90653
28	R	302.6133	209.1117	1.14123	5.787662	21.20402	79.74849	0.04791448	4.218234	19.03281
29	R	309.9426	214.3028	1.154421	6.304025	21.01287	81.72318	0.04851774	4.161478	19.99151
30	R	290.5137	190.7703	1.180543	8.424597	21.29008	74.76164	0.03481699	3.798329	22.14147
31	R	264.779	185.0329	1.236705	7.477044	19.3007	68.14463	0.03559718	3.257943	19.9288
32	R	182.0679	160.8791	2.305367	8.770056	7.366035	58.25157	0.02915993	1.333791	15.91537
33	R	96.21085	93.44856	2.22745	11.01053	7.487364	27.00022	0.01033183	0.04885793	13.94024
34	R	-2.147858	35.87912	2.469797	10.6254	4.540978	0.1539001	0.0005589165	-1.001874	9.623546
35	R	206.3116	147.2226	1.562132	11.57878	13.35509	53.27651	0.01707208	1.258276	20.56206
36	R	187.5558	142.1416	1.366006	8.101585	14.48057	47.51697	0.02427626	1.602733	16.79317
37	R	207.4284	168.7206	2.424866	8.000704	5.559812	73.77534	0.03150713	2.286184	18.63691
38	R	149.4681	122.2414	0.8006789	3.743309	13.31525	37.82116	0.02616863	1.262584	12.3272
39	R	128.4528	100.6417	0.7373622	2.218628	11.70005	30.20584	0.02663895	1.3029	7.704613
40	R	360.3361	267.6172	2.980327	8.420826	3.461347	136.6642	0.05230123	4.966616	29.81557
41	R	67.34041	17.54257	1.422841	-0.1563721	3.165001	25.86943	0.006327856	2.397934	1.111376
42	R	61.27768	30.8145	1.169064	0.7597351	6.217396	15.94258	0.0101852	1.458471	0.8326283
43	R	163.3018	125.65	1.559342	4.88533	7.432486	54.47084	0.02464203	1.911374	14.07451

Ready NUM

Rysunek 3.21. Wyniki prognoz wskaźników fizyko-chemicznych na podstawie współrzędnych punktu monitoringowego uzyskane z modelu RBF dla pliku roboczego (*run*)



Rysunek 3.22. Prognozowanie stężeń cynku [mg/dm^3] na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB, C, D

Najmniejszy błąd prognoz cechuje sumę substancji rozpuszczonych −0.12% i odczyn pH 0.31%. W przypadku fluorków średni błąd względny prognoz osiąga wartość −50%, a przypadku krzemionki wartość maksymalną 108.83%.

Rozkłady błędów względnych prognoz charakteryzuje w przypadku niektórych zmiennych bardzo duży rozrzut — np. dla wapnia −6500–7500%, czy w przypadku krzemionki −1000–19000%. Rozrzut ten spowodowany jest zwykle przez jeden, dwa odbiegające wyniki prognoz.

W części przypadków średni błąd względny prognoz ma wartość ujemną, co świadczy o tendencji do zawyżania wartości prognozowanych w stosunku do wartości prawdziwych.

Tabela 3.4. Analiza prognoz wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór roboczy, punkty klasy AB, C, D (16 zmiennych docelowych)

Lp.	Wskaźnik jakości wód	Średni błąd względny prognoz *B*	Współcz. korelacji
1.	Temperatura	0.89%	0.500
2.	Odczyn pH	0.31%	0.531
3.	Suma substancji rozp.	−0.12%	0.685
4.	Zasadowość ogólna	−4.86%	0.762
5.	Twardość ogólna	−19.55%	0.683
6.	Sód	−6.59%	0.543
7.	Magnez	−2.01%	0.455
8.	Wapń	4.46%	0.689
9.	Chlorki	11.06%	0.479
10.	Siarczany	−0.96%	0.418
11.	Krzemionka	108.83%	0.707
12.	Fluorki	−50.00%	0.750
13.	Cynk	3.63%	0.383
14.	Współczynnik absorpcji UV	10.19%	0.529
15.	Rozpuszczony węgiel organiczny	1.78%	0.494
16.	Utlenialność ChZT-Mn	0.65%	0.579

Współczynniki korelacji wartości obserwowanych z prognozowanymi mieszczą się w zakresie 0.383–0.762. Nie są to więc prognozy dobrej jakości.

Nie stwierdzono związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników (tab. 1.13). Poziom tych błędów zależy jedynie od konfiguracji sieci.

3.1.2. Prognozy dla punktów RMWP o klasie zagrożenia AB (wariant 2.)

Kolejny zbiór *zbior02.sav*, przygotowany zgodnie z wariantem 2. składa się wyłącznie z punktów klasy AB (tab. 3.1).

Takie ograniczenie pozwoli ocenić, czy na jakość uzyskiwanych prognoz ma wpływ zawężenie wejściowego zbioru danych do punktów najbardziej zagrożonych (punktów RMWP klasy AB).

Zbiór danych wejściowy podzielono na podzbiory: treningowy, testowy i walidacyjny. Podzbiór treningowy obejmuje 80% obserwacji (121 obserwacji), w podzbiorach walidacyjnym i testowym jest po 10% obserwacji (15 obserwacji).

Analizie poddano trzy modele sieci: MLP, RBF i Bayesa, podobnie jak w przypadku wariantu 1. Konfiguracje tych modeli i uzyskane dla zbioru trenigowego wyniki prognoz wskaźników fizyko-chemicznych wód zestawione są w tabelach 3.5, 3.6.

3.1.2.1. Sieć typu MLP

Charakterystyka struktur sieci zestawionych w tabeli 3.5:

- *mlp12.nno* — struktura sieci: 2–12–16 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp13.nno* — struktura sieci: 2–12–16; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp14.nno* — struktura sieci: 2–12–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania

Tabela 3.5. Błędy względne prognoz MA% (liczone wg wzoru (3.3)) wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu MLP, zbiór treningowy, punkty klasy AB (16 zmiennych docelowych)

		Wartość błędu prognozy dla wskaźnika jakości wód [%]															
Lp.	Zbiór wynikowy	Temperatura	Odczyn pH	Suma subst. rozp.	Zasadowość ogólna	Twardość ogólna	Sód	Magnez	Wapń	Chlorki	Siarczany	Krzemionka	Fluorki	Cynk	Współcz. absorpcji UV	Rozpuszczony węgiel org.	Utlenialność ChZT-Mn
1.	*mlp12.nno*	8.93	3.54	26.71	22.79	23.92	49.77	47.83	25.12	49.35	50.30	23.87	33.26	55.46	63.98	32.31	31.57
2.	*mlp13.nno*	9.86	3.81	27.31	23.40	25.01	53.45	50.22	25.79	51.15	52.02	27.60	39.55	57.89	65.67	34.38	34.22
3.	*mlp14.nno*	10.43	4.03	29.05	30.19	28.43	56.22	52.58	28.29	56.13	53.27	39.70	44.47	64.49	64.75	37.64	33.09
4.	*mlp15.nno*	10.19	4.16	28.29	27.02	25.86	54.14	51.44	26.57	49.79	52.40	29.06	39.48	60.15	66.13	35.74	36.53
5.	*mlp16.nno*	10.44	4.05	29.07	30.29	29.15	61.07	55.30	28.56	53.83	56.09	42.58	45.93	62.61	67.85	38.88	35.37
6.	*mlp17.nno*	10.08	3.72	25.58	22.51	24.21	53.25	50.76	25.67	48.17	49.39	27.84	34.72	60.86	62.75	33.13	31.34
7.	*mlp18.nno*	9.71	3.69	26.99	21.74	23.49	52.21	49.07	25.66	49.40	51.25	26.22	35.05	58.94	63.69	33.84	35.14
8.	*mlp19.nno*	9.83	3.85	27.00	26.22	26.87	52.59	50.01	28.18	49.72	50.53	28.88	35.95	58.80	64.35	33.41	32.79
9.	*mlp20.nno*	9.36	3.66	27.42	21.02	24.47	50.01	48.39	25.87	49.05	50.06	24.00	32.92	56.33	64.32	32.64	33.32
10.	*mlp21.nno*	9.43	3.54	27.57	22.64	24.93	52.86	49.39	24.96	50.96	53.83	27.66	35.35	56.98	63.02	34.89	32.19
11.	*mlp22.nno*	10.33	4.31	30.20	29.91	28.98	58.99	53.61	29.18	53.72	56.11	41.12	44.70	62.41	62.69	40.21	33.86

wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp15.nno* — konfiguracja jak w przypadku *mlp14.nno*, zmieniona została metoda opcja uaktualniania wag — wagi są uaktualniane po każdym kolejnym etapie (*pattern*);
- *mlp16.nno* – struktura sieci: 2–12–16; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp17.nno* — struktura sieci: 2–12–12–16 (dołożona została druga warstwa ukryta z 12 neuronami); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); zmiana ta spowodowała znaczne wydłużenie czasu „uczenia się" sieci, bez poprawy jakości uzyskanych wyników;
- *mlp18.nno* — struktura sieci: 2–12–12–16; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp19.nno* — struktura sieci: 2–12–12–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp20.nno* — struktura sieci: 2–16–16 (sieć z jedną warstwą ukrytą, zmiana liczby neuronów w tej warstwie); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); w tym przypadku obserwuje się znacznie dłuższy czas uczenia sieci;
- *mlp21.nno* — struktura sieci: 2–32–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp22.nno* — struktura sieci: 2–12–16; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: normalny (*gaussian*).

Średnie błędy względne prognoz w przypadku zbioru obejmującego punkty o klasie zagrożenia AB kształtują się na poziomie od kilku do kilkudziesięciu procent, są jednak mniejsze niż dla pełnego zbioru danych (wariant 1.), w skład którego wchodziły punkty reprezentujące wszystkie klasy zagrożenia wód: AB, C, D.

Najlepsze wyniki (najmniejsze błędy względne prognoz) dla sieci MLP uzyskano przy domyślnej konfiguracji modułu (plik wynikowy *mlp12.nno*).

3.1.2.2. Sieć typu RBF

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.6:

- *rbf16.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf17.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf18.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf19.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf20.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.3; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf21.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.5; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);

Tabela 3.6. Błędy względne prognoz MA% (liczone wg wzoru (3.3)) wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór treningowy, punkty klasy AB (16 zmiennych docelowych)

		Wartość błędu prognozy dla wskaźnika jakości wód [%]																
Lp.	Zbiór wynikowy	Temperatura	Odczyn pH	Suma subst. rozp.	Zasadowość ogólna	Twardość ogólna	Sód	Magnez	Wapń	Chlorki	Siarczany	Krzemionka	Fluorki	Cynk	Współcz. absorpcji UV	Rozpuszczony węgiel org.	Utlenialność ChZT-Mn
1.	*rbf16.nno*	8.23	3.25	25.57	20.15	22.92	48.73	42.13	23.69	46.94	46.18	18.44	28.30	53.47	61.10	32.12	31.44
2.	*rbf17.nno*	10.47	4.45	30.48	30.87	31.33	63.29	54.73	30.81	56.47	56.92	38.69	43.56	59.96	70.32	39.22	36.82
3.	*rbf18.nno*	8.26	3.47	26.69	23.35	25.84	55.41	42.09	25.66	41.82	43.90	27.81	34.60	50.87	59.52	30.17	30.51
4.	*rbf19.nno*	8.41	3.28	26.47	21.15	24.11	52.95	41.31	24.93	44.62	44.40	23.70	30.32	51.68	60.27	30.84	31.22
5.	*rbf20.nno*	9.34	3.69	27.89	23.45	24.94	56.07	48.88	25.19	48.05	48.11	29.29	34.09	54.69	62.86	33.59	32.43
6.	*rbf21.nno*	9.26	3.59	27.63	22.90	24.73	55.49	47.53	25.79	48.41	48.11	25.91	34.12	53.69	62.33	33.45	31.88
7.	*rbf22.nno*	10.18	4.08	30.82	30.93	31.47	62.93	50.74	30.49	52.79	55.66	35.07	42.60	57.89	66.73	36.75	35.91
8.	*rbf23.nno*	8.89	3.99	27.25	26.48	26.01	51.35	48.21	27.75	45.64	52.47	39.03	42.25	52.37	57.17	34.52	30.21
9.	*rbf24.nno*	10.37	4.43	35.69	38.59	36.98	65.63	56.46	37.69	60.15	59.34	44.05	51.66	64.73	68.40	39.34	38.32
10.	*rbf25.nno*	9.25	3.49	26.21	22.53	23.87	48.83	46.75	25.10	46.01	49.77	23.04	35.17	56.49	63.20	32.18	31.05
11.	*rbf26.nno*	10.16	4.28	30.42	31.32	31.24	64.64	55.33	30.39	55.71	57.09	39.17	42.78	59.12	70.16	39.44	37.34
12.	*rbf27.nno*	10.39	4.59	35.62	38.84	36.81	65.99	56.57	37.52	60.35	59.23	45.06	52.11	65.59	67.83	38.79	38.29
13.	*rbf28.nno*	10.04	4.12	30.72	30.30	30.76	62.32	50.16	30.81	51.54	54.31	34.45	43.25	59.73	66.51	36.32	35.55
14.	*rbf29.nno*	8.01	3.86	25.42	25.54	24.41	47.69	45.57	26.17	44.24	50.83	38.47	39.94	51.67	55.13	32.06	27.64
15.	*rbf30.nno*	10.15	4.15	31.96	29.58	30.64	63.66	52.89	31.72	58.31	59.86	43.59	46.95	59.51	69.41	37.43	36.15

- *rbf22.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf23.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf24.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf25.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf26.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf27.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf28.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf29.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf30.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: próbny (*Trial*).

Wyniki prognoz wskaźników fizyko-chemicznych wód uzyskane dla sieci typu RBF kształtują się na poziomie kilku–kilkudziesięciu procent, są jednak lepsze niż w przypadku sieci o topologii MLP, ponadto średnie błędy względne prognoz dla punktów o klasie zagrożenia AB są niższe niż w przypadku pełnego zbioru punktów RMWP (wariant 1., trzy klasy zagrożenia wód).

Najmniejsze błędy względne prognoz dla pliku z wariantu 2. uzyskano przy domyślnej konfiguracji sieci RBF.

3.1.2.3. Sieć typu Bayesa

Podobnie jak w przypadku zbioru z wariantu 1. (*zbior01.sav*), nie udało się uzyskać wyników prognoz dla tego typu sieci, gdyż na pewnym etapie treningu program „zawieszał się", niezależnie od zmiany parametrów sieci — błąd treningu cały czas bardzo szybko wzrastał.

3.1.2.4. Wybór najlepszego modelu sieci

Po ograniczeniu obserwacji w pliku wejściowym do punktów o klasie zagrożenia AB uzyskano lepsze prognozy niż dla zbioru obejmującego obserwacje reprezentujące wszystkie klasy zagrożenia wód (tab. 3.2, 3.3).

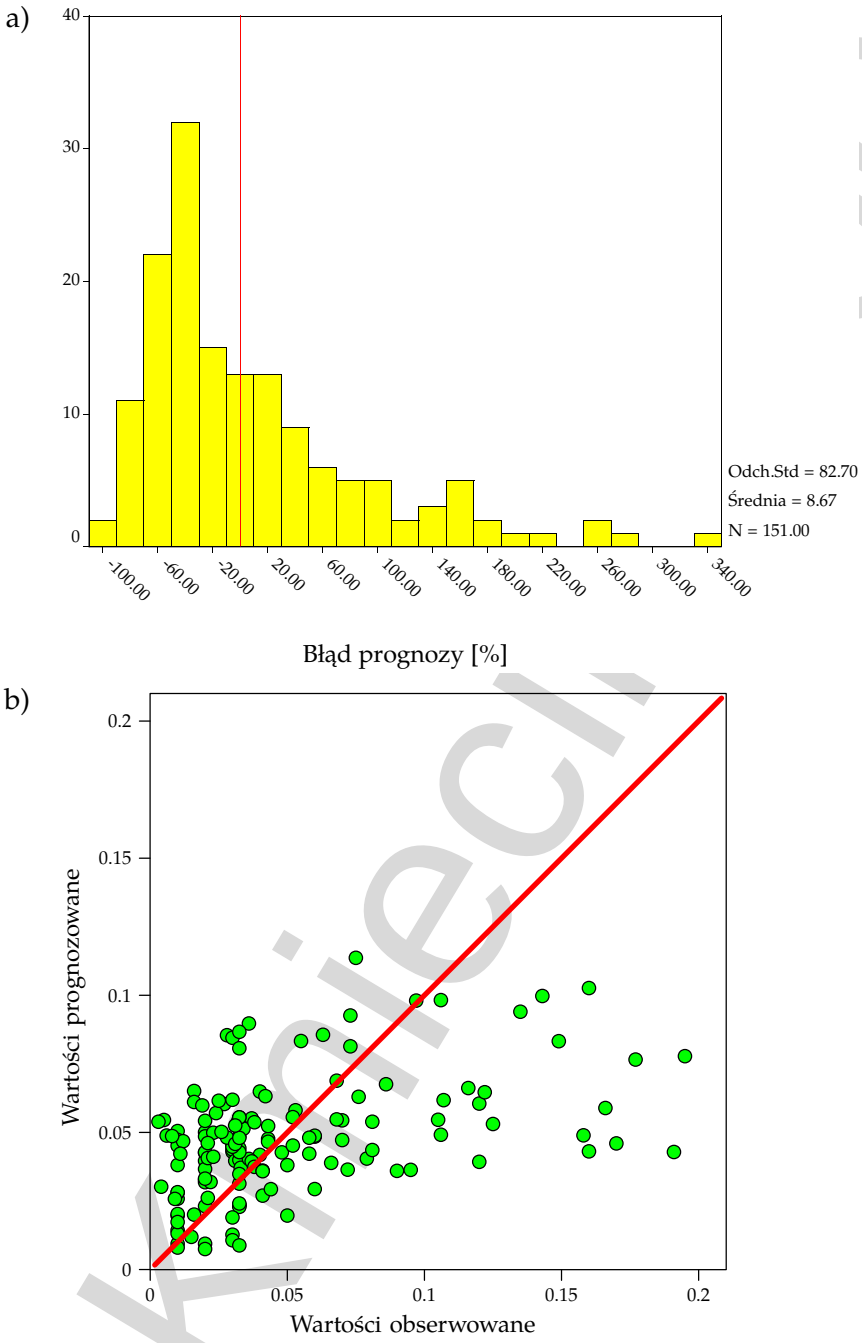
Najlepsze wyniki — najmniejszy średni błąd względny prognoz uzyskano dla sieci typu RBF (tab. 3.5, 3.6) przy domyślnej konfiguracji modułu (odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: *Sample*).

W celu sprawdzenia zdolności prognozowania „nauczonego" modelu, do struktury wprowadzono dane zewnętrzne, plik roboczy (*run*) *run02.sav* z tymi samymi danymi, na których sieć się uczyła. Wyniki prognoz dla pliku roboczego zostały ponownie zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS (*output2.sav*) i zostały obliczone błędy względne prognoz B:

$$B = \frac{x_{obs} - x_{progn}}{x_{obs}} \cdot 100\% \tag{3.5}$$

gdzie: x_{obs} — wartość obserwowana; x_{progn} — wartość prognozowana.

Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych. Na rysunku 3.23 przedstawiono histogram rozkładu błędów i wykres rozrzutu wartości obserwowanych i prognozowanych dla cynku.



Rysunek 3.23. Prognozowanie stężeń cynku [mg/dm^3] na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB

Pełny raport z tej części analizy znajduje się na płycie CD-ROM dołączonej do pracy, w pliku *output2.spo*. W tabeli 3.7 zestawione są średnie błędy prognoz B (obliczone wg wzoru (3.5)) dla analizowanych zmiennych oraz współczynniki korelacji wartości obserwowanych i prognozowanych.

W przypadku cynku, który charakteryzował się najniższą precyzją (tab. 1.13) średni błąd względny prognoz B ma wartość 8.67%, jednak obserwuje się duży rozrzut tych błędów w zakresie od -100 do 400%. Współczynnik korelacji wartości obserwowanych i prognozowanych wynosi zaledwie 0.455.

Tabela 3.7. Analiza prognoz wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór roboczy, punkty klasy AB (16 zmiennych docelowych)

Lp.	Wskaźnik jakości wód	Średni błąd względny prognoz B	Współcz. korelacji
1.	Temperatura	0.89%	0.500
2.	Odczyn pH	0.31%	0.531
3.	Suma subst. rozp.	−0.81%	0.669
4.	Zasadowość ogólna	0.83%	0.808
5.	Twardość ogólna	1.19%	0.702
6.	Sód	6.38%	0.599
7.	Magnez	20.56%	0.564
8.	Wapń	−0.51%	0.731
9.	Chlorki	3.06%	0.555
10.	Siarczany	−1.99%	0.544
11.	Krzemionka	2.17%	0.904
12.	Fluorki	−10.86%	0.837
13.	Cynk	8.67%	0.455
14.	Współczynnik absorpcji UV	10.56%	0.434
15.	Rozpuszczony węgiel organiczny	2.14%	0.622
16.	Utlenialność ChZT-Mn	−0.41%	0.560

Średni błąd względny prognoz B wskaźników fizyko-chemicznych wód kształtuje się na niskim poziomie, od setnych części procenta do kilkunastu procent. Najmniejszy błąd cechuje prognozy odczynu pH −0.04%, a największy — prognozy oznaczeń magnezu 20.56%.

Rozkład błędów względnych prognoz charakteryzuje się w przypadku niektórych zmiennych dużym rozrzutem — np. dla magnezu od −100 do 1600%. Rozrzut ten jest jednak znacznie mniejszy niż w przypadku prognoz dla zbioru z wariantu 1. analizy (punkty reprezentujące wszystkie klasy zagrożenia wód).

W części przypadków średni błąd względny prognoz ma wartość ujemną, co świadczy o tendencji do zawyżania wartości prognozowanych w stosunku do wartości prawdziwych.

Współczynniki korelacji wartości obserwowanych z prognozowanymi mieszczą się w zakresie 0.434–0.904, co oznacza, że ograniczenie zbioru danych wejściowych do obserwacji o klasie zagrożenia AB korzystnie wpłynęło na jakość uzyskanych prognoz.

Również w tym przypadku nie stwierdzono związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników (tab. 1.13). Poziom tych błędów zależy jedynie od konfiguracji sieci.

3.1.3. Prognozy dla punktów RMWP o klasie zagrożenia AB z ograniczoną liczbą zmiennych (wariant 3.)

W celu sprawdzenia, czy gorsza jakość prognoz nie ma swojej przyczyny w sporej liczbie braków danych zastępowanych medianą (tab. 1.13), przygotowano kolejny plik testowy (*zbior03.sav*) według wariantu 3. (tab. 3.1), ograniczony do sześciu zmiennych docelowych (wskaźników fizyko-chemicznych).

Zbiór ten, po wczytaniu do programu Neural Connection, podzielono na podzbiory: treningowy — obejmujący 80% obserwacji (114 obserwacji), walidacyjny 10% (15 obserwacji) i testowy 10% (14 obserwacji).

Przy dwóch zmiennych typu wejściowego ($M = 2$) i sześciu zmiennych docelowych ($N = 6$) w zbiorze treningowym powinno znaleźć się co najmniej $10(M + N) = 10(2 + 6) = 80$ obserwacji (SPSS, 1997). W przypadku badanego zbioru warunek ten został spełniony.

Następnie testowano modele sieci o różnych parametrach, podobnie jak w przypadku poprzednich wariantów (str. 101, 113). Konfiguracje poszczególnych modeli i uzyskane dla zbioru trenigowego wyniki prognoz wskaźników chemicznych wód zestawione są w tabelach 3.8, 3.9.

3.1.3.1. Sieć typu MLP

Charakterystyka struktur sieci zestawionych w tabeli 3.8:

- *mlp23.nno* — struktura sieci: 2–4–6 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp24.nno* — struktura sieci: 2–4–6; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp25.nno* — struktura sieci: 2–4–6; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp26.nno* — konfiguracja jak w przypadku *mlp25.nno*, zmieniona została metoda opcja uaktualniania wag – wagi są uaktualniane po każdym kolejnym etapie (*pattern*);
- *mlp27.nno* — struktura sieci: 2–4–6; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp28.nno* — struktura sieci: 2–4–4–6 (dołożona została druga warstwa ukryta z 4 neuronami); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); zmiana ta spowodowała znaczne wydłużenie czasu „uczenia się" sieci, bez poprawy jakości uzyskanych wyników;
- *mlp29.nno* — struktura sieci: 2–4–4–6; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*);

Tabela 3.8. Błędy względne prognoz MA% (liczone wg wzoru (3.3)) wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu MLP, zbiór treningowy, punkty klasy AB (6 zmiennych docelowych)

Lp.	Zbiór wynikowy	Wartość błędu prognozy [%]					
		Suma subst. rozp.	Zasadowość ogólna	Twardość ogólna	Magnez	Wapń	Krzemionka
1.	*mlp23.nno*	31.89	30.65	29.93	52.62	30.34	31.90
2.	*mlp24.nno*	36.99	38.09	37.32	62.23	35.88	45.93
3.	*mlp25.nno*	31.92	31.83	29.99	57.81	31.85	39.10
4.	*mlp26.nno*	33.37	33.57	32.09	58.88	33.46	39.91
5.	*mlp27.nno*	31.91	32.34	31.13	58.06	32.99	37.22
6.	*mlp28.nno*	37.30	37.39	36.18	67.94	39.90	49.79
7.	*mlp29.nno*	38.65	41.54	38.96	61.64	41.06	53.46
8.	*mlp30.nno*	34.19	35.75	34.76	54.00	37.03	37.69
9.	*mlp31.nno*	29.23	27.14	27.15	51.55	29.37	34.67
10.	*mlp32.nno*	30.93	30.40	30.44	56.05	31.51	38.79
11.	*mlp33.nno*	32.28	34.16	32.16	58.22	34.34	41.62

- *mlp30.nno* — struktura sieci: 2–4–4–6; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp31.nno* — struktura sieci: 2–6–6 (sieć z jedną warstwą ukrytą, zmiana liczby neuronów w tej warstwie); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*), w tym przypadku można zaobserwować znacznie dłuższy czas uczenia się sieci;
- *mlp32.nno* — struktura sieci: 2–12–6; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp33.nno* — struktura sieci: 2–4–6; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: normalny (*gaussian*).

Średnie błędy względne prognoz w przypadku zbioru obejmującego punkty RMWP o klasie zagrożenia AB przy ograniczonej liczbie zmiennych kształtują się na poziomie kilkudziesięciu procent (20–70%), są większe niż dla zbiorów danych analizowanych w poprzednich wariantach.

Najlepsze wyniki (najmniejsze błędy względne prognoz MA%) uzyskano dla sieci MLP o konfiguracji 2–6–6, dla której wyniki zapisane są w pliku *mlp31.nno*.

3.1.3.2. Sieć typu RBF

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.9:

- *rbf31.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf32.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf33.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf34.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf35.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.3; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf36.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.5; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf37.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf38.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf39.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf40.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf41.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf42.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: próbny (*Trial*);
- *rbf43.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: losowy (*Random*);
- *rbf44.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: próbny (*Trial*);

- *rbf45.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: próbny (*Trial*).

Tabela 3.9. Błędy względne prognoz MA% (liczone wg wzoru (3.3)) wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór treningowy, punkty klasy AB (6 zmiennych docelowych)

Lp.	Zbiór wynikowy	Wartość błędu prognozy [%]					
		Suma subst. rozp.	Zasadowość ogólna	Twardość ogólna	Magnez	Wapń	Krzemionka
1.	*rbf31.nno*	26.87	21.63	24.16	46.93	25.02	27.43
2.	*rbf32.nno*	31.25	32.58	31.25	51.67	32.37	37.56
3.	*rbf33.nno*	23.04	20.21	21.47	36.55	22.60	23.38
4.	*rbf34.nno*	29.53	24.92	27.36	47.77	28.36	24.81
5.	*rbf35.nno*	28.51	23.23	25.86	45.94	27.55	21.53
6.	*rbf36.nno*	26.65	20.44	23.96	43.89	25.21	19.89
7.	*rbf37.nno*	33.24	32.84	33.83	53.02	34.67	41.14
8.	*rbf38.nno*	32.52	30.83	30.96	56.83	34.03	43.95
9.	*rbf39.nno*	36.82	40.49	37.52	62.11	37.89	52.56
10.	*rbf40.nno*	31.74	31.06	29.61	53.27	32.17	32.73
11.	*rbf41.nno*	30.86	27.85	29.30	48.07	30.17	23.83
12.	*rbf42.nno*	37.15	40.87	37.94	61.99	38.49	50.52
13.	*rbf43.nno*	35.62	35.71	35.99	56.40	38.51	46.68
14.	*rbf44.nno*	24.56	22.41	22.19	35.69	22.32	25.74
15.	*rbf45.nno*	33.77	32.69	32.06	59.33	34.95	43.39

Sieć typu RBF daje lepsze wyniki prognoz niż sieć MLP (mniejszy średni błąd względny prognoz MA%) i znacznie szybciej „uczy się".

Najlepsze wyniki (najmniejsze błędy względne prognoz MA%) uzyskano dla sieci RBF, dla której wyniki zapisane są w pliku *rbf44.nno*.

3.1.3.3. Sieć typu Bayesa

Ponieważ sieć Bayesa nie korzysta ze zbioru walidacyjnego, dokonano podziału zbioru danych wejściowych na podzbiory treningowy i testowy w proporcjach 90% : 10% (129 : 14 obserwacji).

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.10[(4)]:

- *bayes01.nno* — struktura sieci: 2–4–6 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Single Weight Group*; wybór modelu: *Commitee Decision*;
- *bayes02.nno* — struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Single Weight Group*; wybór modelu: *Most Likely Model*;
- *bayes03.nno* — struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weights Grouped by Layer*; wybór modelu: *Most Likely Model*;
- *bayes04.nno* — struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Separate Weights and Biases*; wybór modelu: *Most Likely Model*;

(4) Opis poszczególnych opcji można znaleźć w dokumentacji do programu Neural Connection (SPSS, 1997) oraz w dodatku B niniejszej pracy.

- *bayes05.nno* — struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Auto Relevance Detection*; wybór modelu: *Most Likely Model*;
- *bayes06.nno* — struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weights Grouped by Layer*; wybór modelu: *Commitee Decision*;
- *bayes07.nno* — struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Separate Weights and Biases*; wybór modelu: *Commitee Decision*;
- *bayes08.nno* — struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Auto Relevance Detection*; wybór modelu: *Commitee Decision*.

Tabela 3.10. Błędy względne prognoz MA% (liczone wg wzoru (3.3)) wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu Bayesa, zbiór treningowy, punkty klasy AB (6 zmiennych docelowych)

Lp.	Zbiór wynikowy	Wartość błędu prognozy [%]					
		Suma subst. rozp.	Zasadowość ogólna	Twardość ogólna	Magnez	Wapń	Krzemionka
1.	*bayes01.nno*	35.34	37.63	36.15	59.36	36.25	46.07
2.	*bayes02.nno*	36.17	36.51	35.41	60.81	34.92	43.32
3.	*bayes03.nno*	35.14	37.72	35.44	61.32	36.10	44.73
4.	*bayes04.nno*	32.18	31.66	30.51	54.99	32.56	39.20
5.	*bayes05.nno*	35.08	35.02	35.38	58.45	34.62	42.13
6.	*bayes06.nno*	35.14	37.72	35.44	61.32	36.10	44.73
7.	*bayes07.nno*	32.18	31.66	30.51	54.99	32.56	39.19
8.	*bayes08.nno*	35.08	35.02	35.38	58.45	34.62	42.13

Najlepsze wyniki (najmniejsze błędy względne prognoz MA%) uzyskano dla sieci Bayesa o strukturze 2–4–6, dla której wyniki zapisane są w pliku *bayes04.nno*.

Sieć ta daje gorsze wyniki prognoz (wyższe błędy względne prognoz MA%) niż sieci MLP i RBF, na poziomie 30–60%.

3.1.3.4. Wybór najlepszego modelu sieci

Najlepsze wyniki — najmniejszy średni błąd względny prognoz siedmiu wskaźników jakości wód uzyskano ponownie dla sieci typu RBF (tab. 3.8, 3.9, 3.10) przy konfiguracji modułu: odległość — *Euclidean*; funkcja nieliniowa — *Inv. Quadratic*; parametr funkcji — 0.5; liczba centrów: 5; rozkład centrów — *Sample*.

W celu sprawdzenia zdolności prognozowania „nauczonego" modelu, do struktury wprowadzono dane zewnętrzne, plik roboczy (*run*) *run03.sav* z tymi samymi danymi, na których sieć się uczyła.

Wyniki prognoz dla pliku roboczego zostały ponownie zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS (*output3.sav*). W programie SPSS obliczono błędy względne prognoz B:

$$B = \frac{x_{obs} - x_{progn}}{x_{obs}} \cdot 100\% \tag{3.6}$$

gdzie: x_{obs} — wartość obserwowana; x_{progn} — wartość prognozowana.

Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych. Pełny raport z tej części analizy znajduje się na płycie CD-ROM dołączonej do pracy, w pliku *output3.spo*.

W tabeli 3.11 zestawione są średnie błędy prognoz B (obliczone wg wzoru (3.6)) dla analizowanych zmiennych oraz współczynniki korelacji wartości obserwowanych i prognozowanych.

Tabela 3.11. Analiza prognoz wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór roboczy, punkty klasy AB (6 zmiennych docelowych)

Lp.	Wskaźnik jakości wód	Średni błąd względny prognoz B	Współcz. korelacji
1.	Suma subst. rozp.	−0.22%	0.671
2.	Zasadowość ogólna	1.03%	0.816
3.	Twardość ogólna	3.79%	0.724
4.	Magnez	12.42%	0.634
5.	Wapń	3.29%	0.730
6.	Krzemionka	4.12%	0.895

Średni błąd względny prognoz kształtuje się na poziomie rzędu kilku procent. Najmniejszy błąd cechuje prognozy sumy substancji rozpuszczonych −0.22%, a największy — podobnie jak w przypadku zbioru z 16 zmiennymi docelowymi — prognozy oznaczeń magnezu 12.42% (ma on jednak dwukrotnie mniejszą wartość niż dla zbioru z 16 zmiennymi docelowymi). Rozkład błędów względnych prognoz magnezu charakteryzuje się ponadto dużym rozrzutem −400–1100%. W jednym przypadku (suma substancji rozpuszczonych) średni błąd względny prognoz ma wartość ujemną, co świadczy o tendencji do zawyżania prognoz w stosunku do wartości prawdziwych.

Współczynniki korelacji wartości obserwowanych z prognozowanymi mieszczą się w zakresie 0.594–0.895. Ograniczenie zbioru danych wejściowych do 6 zmiennych docelowych (klasa zagrożenia AB) nie wpłynęło zatem w istotny sposób na jakość uzyskiwanych prognoz.

Podobnie jak w przypadku wariantów 1. i 2., również i w tym przypadku nie stwierdzono związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników (tab. 1.13). Poziom błędów prognoz zależy jedynie od konfiguracji sieci.

3.1.4. Optymalizacja gęstości opróbowania

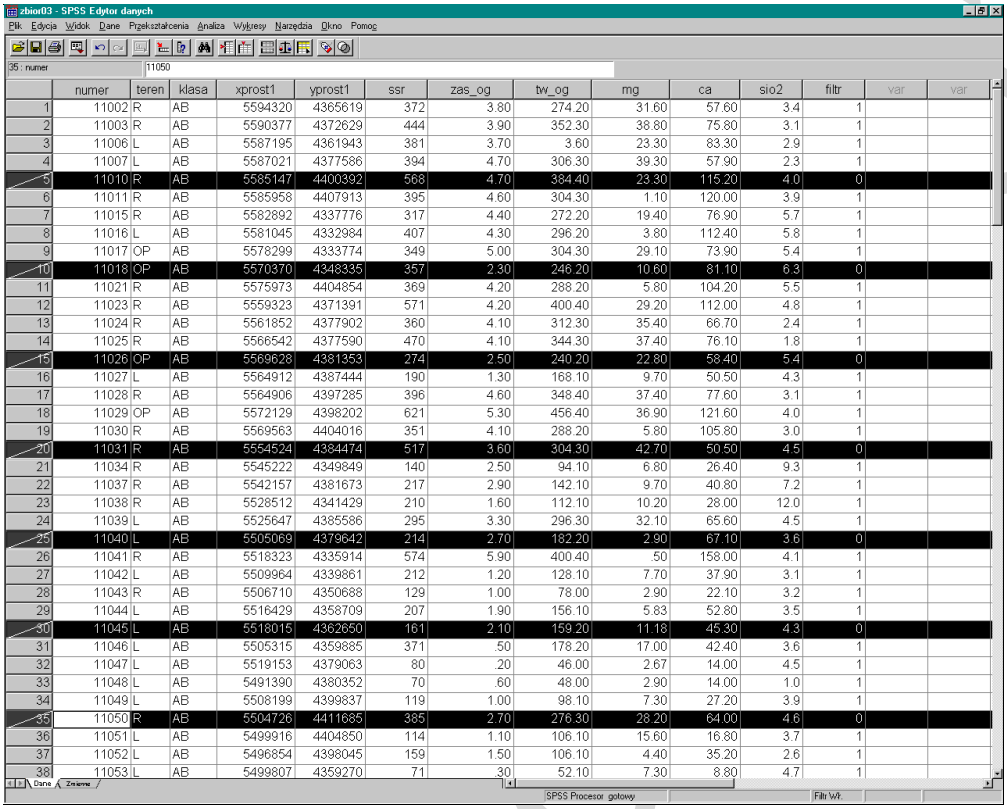
Aby sprawdzić, czy po wyłączeniu z istniejącej sieci pomiarowej części punktów, będzie można na podstawie pozostałych — opróbowanych — punktów RMWP prognozować zmiany jakości wód podziemnych w całej sieci monitoringowej, dokonano pewnej modyfikacji zbioru utworzonego wg wariantu 3. (*zbior03.sav*).

Wyłączono ze zbioru danych wejściowych część obserwacji, zapisując je w pliku roboczym, służącym do sprawdzenia jakości uzyskanych prognoz. Następnie uczono sieć na podstawie zbioru danych wejściowych o mniejszej liczebności i sprawdzano, jak sieć zachowa się prognozując wskaźniki fizyko-chemiczne w punktach RMWP „teoretycznie" nieopróbowanych, znajdujących się w pliku roboczym.

Następnie poddano weryfikacji uzyskane dla pliku roboczego wyniki prognoz, porównując je z prawdziwymi wynikami analiz.

Wyłączenie z analizy co piątego punktu RMWP (gęstość sieci 1 pkt RMWP/419.7 km^2)

Ze zbioru *zbior03.sav* usunięto, kolejno, co piątą obserwację — punkt RMWP (rys. 3.24) i zapisano go pod nazwą *zbior04.sav* — zbiór liczy 115 obserwacji. Oznacza to, że gęstość opróbowania sieci monitoringowej wynosi w tym przypadku 1 pkt RMWP/419.7 km^2.



Rysunek 3.24. Zbiór *zbior03.sav* przygotowany wg wariantu 3. Szarym kolorem podświetlone są obserwacje, które zostaną zapisane do roboczego pliku danych

Wyłączone z analizy obserwacje (punkty RMWP: 11010, 11018, 11026, 11031, 11040, 11045, 11050, 11055, 21005, 21013, 21020, 21027, 21032, 21037, 21042, 21048, 21056, 21062, 21067, 21072, 21077, 21082, 21087, 21093, 21099, 21105, 21112, 21117 — patrz rys. 3.25) zapisano w roboczym pliku danych *run04.sav*.

Zbiór *zbior04.sav* po wczytaniu do programu Neural Connection podzielono na podzbiory: treningowy — obejmujący 80% obserwacji (91 obserwacji), walidacyjny i testowy — po 10% obserwacji (12 obserwacji).

Przy dwóch zmiennych typu wejściowego i sześciu zmiennych docelowych w zbiorze treningowym powinno być 80 obserwacji, zatem w analizowanym przypadku warunek ten został spełniony.

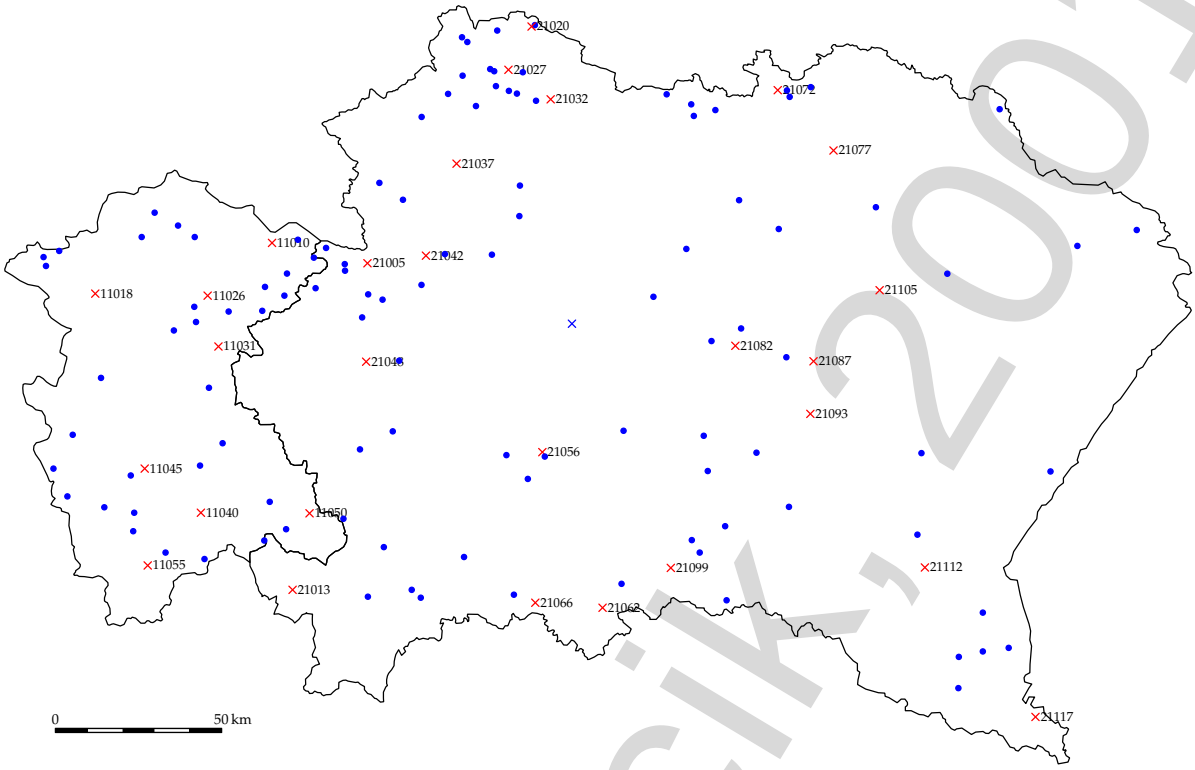
Następnie testowano różne modele sieci. Konfiguracje poszczególnych modeli i uzyskane dla zbioru treningowego wyniki prognoz wskaźników chemicznych wód znajdują się na dołączonej do pracy płycie CD-ROM (pliki *rbf46.nno–rbf59.nno*). Najlepsze wyniki — najmniejszy średni błąd względny prognoz MA% (wzór 3.3) sześciu wskaźników jakości wód uzyskano przy domyślnej konfiguracji sieci typu RBF (plik *rbf45.nno*): odległość — *Euclidean*; funkcja nieliniowa — *Spline*; liczba centrów: 5; rozkład centrów — *Sample*.

W celu sprawdzenia zdolności prognozowania „nauczonego" modelu, do struktury wprowadzono dane zewnętrzne, plik roboczy (*run*) *run04.sav* z danymi dla punktów wyłączonych wcześniej z analizy.

Wyniki prognoz dla pliku roboczego zostały ponownie zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS (*output4.sav*) i obliczono błędy względne prognoz B:

$$B = \frac{x_{obs} - x_{progn}}{x_{obs}} \cdot 100\% \tag{3.7}$$

gdzie: x_{obs} — wartość obserwowana; x_{progn} — wartość prognozowana.



Rysunek 3.25. Regionalny monitoring jakości wód podziemnych (RMWP) w zlewni górnej Wisły. Punkty klasy AB po weryfikacji i usunięciu obserwacji z brakami danych. Obserwacje z plików: • — *zbior04.sav* (plik wejściowy — punkty opróbowane, służące do „uczenia sieci"); × — *run04.sav* (plik roboczy — punkty „teoretycznie nieopróbowane", służące do sprawdzenia poprawności prognoz)

Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych. Pełny raport z tej części analizy znajduje się na płycie CD-ROM dołączonej do pracy, w pliku *output4.spo*.

Tabela 3.12. Analiza prognoz wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór roboczy *run04.sav*, punkty klasy AB wyłączone wcześniej z analizy (6 zmiennych docelowych)

Lp.	Wskaźnik jakości wód	Średni błąd względny prognoz B	Współcz. korelacji
1.	Suma subst. rozp.	3.71%	0.517
2.	Zasadowość ogólna	30.67%	0.503
3.	Twardość ogólna	14.55%	0.581
4.	Magnez	−20.67%	0.204
5.	Wapń	11.93%	0.538
6.	Krzemionka	14.64%	0.785

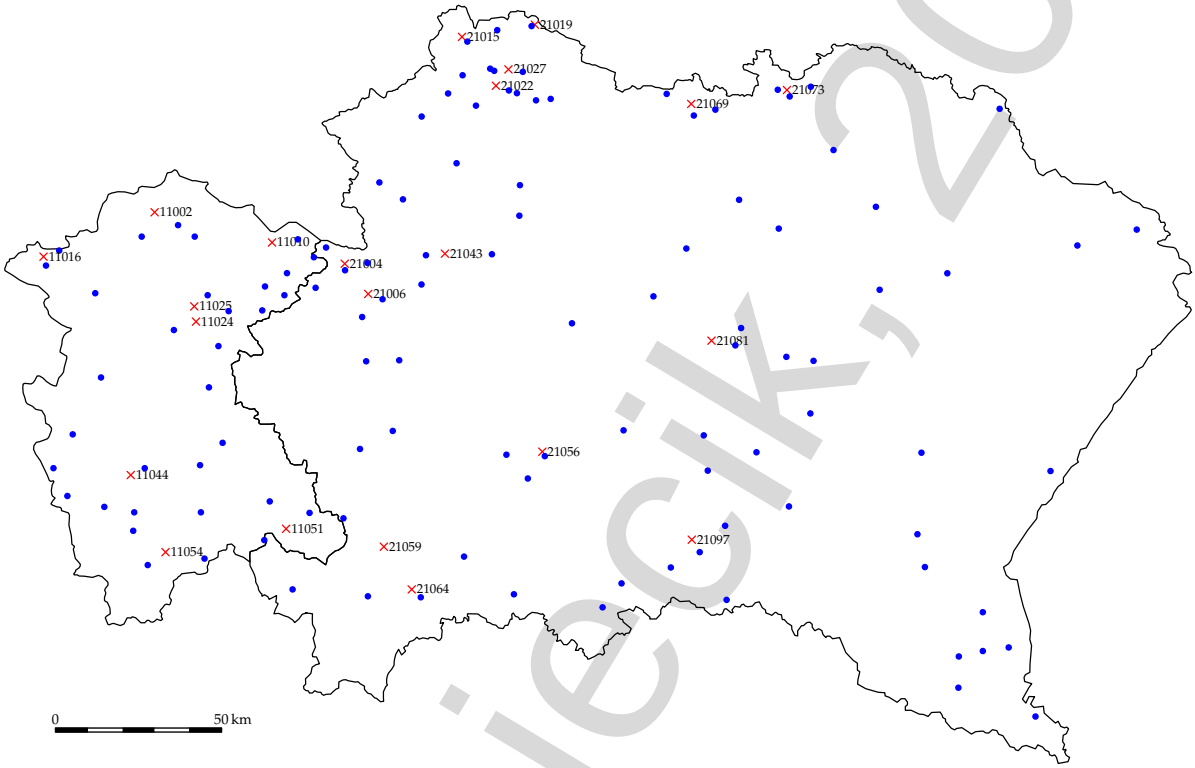
W tabeli 3.12 zestawione są średnie błędy prognoz wskaźników chemicznych oraz współczynniki korelacji wartości obserwowanych i prognozowanych. Średni błąd względny prognoz kształtuje się na poziomie kilku–kilkudziesięciu procent, współczynnik korelacji wartości obserwowanych i prognozowanych zmienia się od 0.204 do 0.785.

Wyłączenie z analizy punktów gęsto położonych (gęstość sieci 1 pkt RMWP/398.9 km^2)

Następnie przyjęto inną strategię wyłączania punktów RMWP z analizy. Z pliku *zbior03.sav* z sześcioma zmiennymi docelowymi (143 obserwacje) wyłączono punkty, w pobliżu których

(w odległości 10–15 km) znajduje się co najmniej dwa inne punkty RMWP, i zapisano go pod nazwą *zbior05.sav* — zbiór liczy 121 obserwacji. Oznacza to, że gęstość opróbowania sieci monitoringowej wynosi w tym przypadku 1 pkt RMWP/398.9 km^2.

Wyłączone z analizy punkty RMWP: 11002, 11010, 11016, 11024, 11025, 11044, 11051, 11054, 21004, 21006, 21015, 21019, 21022, 21027, 21043, 21056, 21059, 21064, 21069, 21073, 21081, 21097 (patrz rys. 3.26) zapisano w roboczym pliku danych *run05.sav*.



Rysunek 3.26. Regionalny monitoring jakości wód podziemnych (RMWP) w zlewni górnej Wisły. Punkty klasy AB po weryfikacji i usunięciu obserwacji z brakami danych. Obserwacje z plików: • — *zbior05.sav* (plik wejściowy — punkty opróbowane, służące do „uczenia sieci"); × — *run05.sav* (plik roboczy — punkty „teoretycznie nieopróbowane", służące do sprawdzenia poprawności prognoz)

Zbiór *zbior05.sav* po wczytaniu do programu Neural Connection podzielono na podzbiory: treningowy — obejmujący 80% obserwacji (97 obserwacji), walidacyjny i testowy — po 10% obserwacji (12 obserwacji).

Następnie testowano różne modele sieci. Konfiguracje poszczególnych modeli i uzyskane dla zbioru treningowego wyniki prognoz wskaźników chemicznych wód znajdują się na dołączonej do pracy płycie CD-ROM (pliki *rbf60.nno–rbf85.nno*). Najlepsze wyniki — najmniejszy średni błąd względny prognoz siedmiu wskaźników jakości wód uzyskano przy następującej konfiguracji sieci (plik *rbf82.nno*): odległość — *Euclidean*; funkcja nieliniowa — *Inv. Quadratic*; parametr funkcji — 0.1; liczba centrów — 5; rozkład centrów — *Sample*.

W celu sprawdzenia zdolności prognozowania „nauczonego" modelu, do struktury wprowadzono dane zewnętrzne, plik roboczy (*run*) *run05.sav* z danymi dla punktów wyłączonych wcześniej z analizy. Wyniki prognoz dla pliku roboczego zostały ponownie zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS (*output5.sav*).

W programie SPSS zostały obliczone błędy względne prognoz B:

$$B = \frac{x_{obs} - x_{progn}}{x_{obs}} \cdot 100\% \tag{3.8}$$

gdzie: x_{obs} — wartość obserwowana; x_{progn} — wartość prognozowana.

Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych. Pełny raport z tej części analizy znajduje się na płycie CD-ROM dołączonej do pracy, w pliku *output5.spo*. W tabeli 3.13 zestawione są średnie błędy prognoz B (obliczone wg wzoru 3.8) dla analizowanych zmiennych oraz współczynniki korelacji wartości obserwowanych i prognozowanych.

Tabela 3.13. Analiza prognoz wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór roboczy *run05.sav*, punkty klasy AB wyłączone wcześniej z analizy (6 zmiennych docelowych)

Lp.	Wskaźnik jakości wód	Średni błąd względny prognoz B	Współcz. korelacji
1.	Suma subst. rozp.	2.79%	0.569
2.	Zasadowość ogólna	−4.06%	0.860
3.	Twardość ogólna	−4.87%	0.749
4.	Magnez	8.89%	0.668
5.	Wapń	−2.81%	0.675
6.	Krzemionka	−14.59%	0.788

Średni błąd względny prognoz we wszystkich przypadkach (za wyjątkiem krzemionki) kształtuje się na poziomie kilku procent, współczynnik korelacji wartości obserwowanych i prognozowanych mieści się w zakresie 0.569–0.860. Oznacza to, że w badanym przypadku, przy opróbowaniu 84% punktów monitoringowych (121 punktów na 143) można z wykorzystaniem sieci neuronowych, na podstawie wyników oznaczeń wód w tych punktach uzyskać wiarygodne prognozy dotyczące jakości wód w punktach nieopróbowanych.

Można zatem ograniczyć liczbę punktów monitoringowych opróbowanych w danej serii pomiarowej np. o ok. 15% — pobrać mniejszą liczbę próbek i na podstawie wiarygodnych oznaczeń stężeń wskaźników chemicznych w tych próbkach uzyskać informacje o jakości wód w całym obszarze dorzecza górnej Wisły. Zmniejszenie liczby pobieranych próbek spowoduje z kolei mniejszy koszt prowadzenia badań monitoringowych.

W obszarze dorzecza górnej Wisły znajduje się 172 punkty RMWP, oznacza to, że 1 punkt RMWP przypada na 280.6 km^2, przy czym w obszarze RZGW Katowice jest to 1 pkt RMWP na 134.2 km^2, a w obszarze RZGW Kraków 1 pkt RMWP na 349.5 km^2. W I serii opróbowania RMWP dorzecza górnej Wisły pobrano 167 próbek, opróbowano średnio 1 punkt RMWP na 289 km^2. W analizowanym przypadku sieć monitoringowa została sztucznie „rozgęszczona" przez wyłączenie z analizy punktów RMWP, w których wystąpiły braki danych.

Przeprowadzone symulacje komputerowe wskazują, że przy gęstości opróbowania 1 pkt RMWP/398.9 km^2 można uzyskać informacje o jakości wód w nieopróbowanych punktach RMWP.

Trudno wyznaczyć tu graniczną wartość, do jakiej można ograniczać liczbę opróbowanych, bądź nieopróbowanych punktów. Należy przeprowadzić co najmniej jedno pełne opróbowanie sieci o największej możliwej gęstości, wykonać pełne analizy wszystkich pobranych próbek i zbudować na tej podstawie model sieci neuronowej. Następnie należy na tym modelu eksperymentować, sprawdzając przy jakiej docelowej gęstości można uzyskiwać wiarygodne prognozy jakości wód w badanym obszarze.

3.2. Klasyfikacja punktu monitoringowego do obszaru o określonym użytkowaniu terenu

Aby stwierdzić czy na podstawie wyników oznaczeń wskaźników jakości wód można uzyskać dane dotyczące sposobu użytkowania terenu w danym punkcie RMWP zbudowano model sieci neuronowej, w której zmiennymi wejściowymi są oznaczenia wskaźników fizykochemicznych jakości wód podziemnych a zmienną docelową — sposób użytkowania terenu.

3.2.1. Prognozy dla punktów RMWP reprezentujących wszystkie klasy zagrożenia wód (wariant 1.)

Przygotowany zgodnie z wariantem pierwszym (tab. 3.1) plik *zbior01a.sav*[5] wczytano wprost do programu Neural Connection, uruchamiając z programu SPSS opcję **Analiza ▶ Neural Connection**.

Następnie dokonano konfiguracji zmiennych (rys. 3.7), w taki sposób, że zmienne: numer identyfikacyjny punktu w bazie MONBADA, klasa zagrożenia wód i współrzędne punktu monitoringowego w układzie 42 zdefiniowano jako zmienne typu opisowego (R), 16 wskaźników fizyko-chemicznych to zmienne wejściowe (I), a zmienną docelową (T) jest sposób użytkowania terenu w otoczeniu punktu RMWP (rys. 3.27). Wskaźniki fizyko-chemiczne charakteryzujące się rozkładem logarytmiczno-normalnym (rozdz. 1.2) poddano, za pomocą narzędzia filtrującego, operacji logarytmowania (SPSS, 1997).

Data Viewer - [Input development data]

File Data Field Window

		Integer R	Symbol T	Symbol R	Float R	Float R	Float I	Float
		NUMER	TEREN	KLASA	XPROST1	YPROST1	TEMP	PH
1	T	11001	L	AB	5596237.0	4354615.0	10.0	
2	V	11002	R	AB	5594320.0	4365619.0	12.0	
3	T	11003	R	AB	5590377.0	4372629.0	9.0	
4	T	11004	L	C	5597748.0	4377335.0	11.0	
5	V	11005	R	D	5592479.0	4384254.0	11.0	
6	X	11006	L	AB	5587195.0	4361943.0	8.0	
7	T	11007	L	AB	5587021.0	4377586.0	9.0	
8	T	11008	R	D	5583416.0	4385205.0	11.0	
9	V	11009	R	D	5587349.0	4392659.0	11.0	
10	X	11010	R	AB	5585147.0	4400392.0	9.0	
11	T	11011	R	AB	5585958.0	4407913.0	9.0	
12	V	11013	OP	AB	5583189.0	4353273.0	11.5	
13	T	11014	OP	AB	5583195.0	4353023.0	12.0	
14	T	11015	R	AB	5582892.0	4337776.0	10.0	
15	T	11016	L	AB	5581045.0	4332984.0	10.0	
16	T	11017	OP	AB	5578299.0	4333774.0	11.0	
17	T	11018	OP	AB	5570370.0	4348335.0	9.5	
18	T	11019	R	AB	5576776.0	4366581.0	12.0	
19	T	11020	R	AB	5576254.0	4385450.0	9.0	
20	T	11021	R	AB	5575973.0	4404854.0	9.0	
21	T	11022	R	AB	5564445.0	4369760.0	11.0	
22	T	11023	R	AB	5559323.0	4371391.0	11.0	
23	X	11024	R	AB	5561852.0	4377902.0	10.0	
24	V	11025	R	AB	5566542.0	4377590.0	10.0	
25	T	11026	OP	AB	5569628.0	4381353.0	9.0	
26	T	11027	L	AB	5564912.0	4387444.0	9.0	
27	T	11028	R	AB	5564906.0	4397285.0	9.0	
28	T	11029	OP	AB	5572129.0	4398202.0	11.0	
29	V	11030	R	AB	5569563.0	4404016.0	8.0	
30	T	11031	R	AB	5554524.0	4384474.0	10.0	

Ready NUM

Rysunek 3.27. Ekran podglądu danych wejściowych. Objaśnienia: R — zmienne typu opisowego; I — zmienne typu wejściowego; T — zmienna docelowa

Zmienna docelowa — sposób użytkowania terenu — przyjmuje trzy wartości: dla obszaru o zagospodarowaniu rolniczym — symbol *R*, w obszarze o zagospodarowaniu leśnym — *L* a w obszarze o zagospodarowaniu osiedlowo-przemysłowym — *OP*.

Następnie, w celu wyboru optymalnej — dającej najlepsze rezultaty prognoz — struktury sieci testowano różne modele z grupy sieci nadzorowanych (*supervised*) — wielowarstwowy perceptron MLP, radialna funkcja bazowa RBF, i sieć Bayesa (podobnie jak w rozdziale 3.1).

3.2.1.1. Sieć typu MLP

Zbiór danych wejściowych został podzielony na podzbiory (rys. 3.12): treningowy (80% wszystkich obserwacji — 133 obserwacje), walidacyjny (10% — 17 obserwacji) i testowy (17 obserwacji).

(5) Ten plik, tak jak i wszystkie pliki wynikowe oraz pliki z modelami omawianych sieci neuronowych znajdują się na płycie CD-ROM dołączonej do niniejszej pracy.

Przy szesnastu zmiennych typu wejściowego ($M = 16$) i jednej zmiennej typu docelowego ($N = 1$) w zbiorze treningowym powinno być co najmniej $10(M + N) = 10(16 + 1) = 170$ obserwacji (SPSS, 1997), zatem w tym przypadku jakość uzyskiwanych prognoz może być nieco gorsza, z uwagi na mniejszą liczbę obserwacji w zbiorze treningowym.

Pierwszą próbę „uczenia" sieci przeprowadzono przy domyślnych ustawieniach opcji modułu MLP. Kolejne modyfikacje dotyczyły algorytmu uczącego i sposobu uaktualniania wag neuronów, dokładano drugą warstwę ukrytą, zmieniano liczbę neuronów w warstwach. Uzyskane dla zbioru trenigowego wyniki prognoz wskaźników chemicznych wód zestawione są w tabeli 3.14.

W tym przypadku parametrem określającym jakość uzyskanych prognoz będzie procent obserwacji poprawnie zaklasyfikowanych (punktów monitoringowych poprawnie przyporządkowanych do obszaru o określonym użytkowaniu terenu).

Tabela 3.14. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu MLP, zbiór treningowy, punkty klasy AB, C, D (16 zmiennych wejściowych, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*mlp01a.nno*	76.12
2.	*mlp02a.nno*	79.10
3.	*mlp03a.nno*	76.12
4.	*mlp04a.nno*	75.37
5.	*mlp05a.nno*	76.12
6.	*mlp06a.nno*	75.37
7.	*mlp07a.nno*	73.13

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.14:

- *mlp01a.nno* — struktura sieci: 16–4–3 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp02a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp03a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp04a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp05a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp06a.nno* — struktura sieci: 16–4–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*);

- *mlp07a.nno* — struktura sieci: 16–4–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*).

Próby zmiany konfiguracji sieci — dokładanie warstw ukrytych, zmiana liczby neuronów w warstwie ukrytej, modyfikacja rozkładu wag neuronów — nie dały pozytywnych rezultatów (poprawy uzyskanych prognoz). Sieć typu MLP poprawnie prognozuje ok. 70% punktów RMWP.

3.2.1.2. Sieć typu RBF

Zachowano podział zbioru danych wejściowych jak w przypadku sieci MLP: podzbiór treningowy (80% wszystkich obserwacji — 133 obserwacje), walidacyjny i testowy (po 10% — 17 obserwacji).

Pierwszą próbę „uczenia" sieci przeprowadzono przy domyślnych ustawieniach modułu. Kolejne modyfikacje dotyczyły odległości (*error distance*), rodzaju funkcji nieliniowej i jej parametrów. Konfiguracje poszczególnych modeli sieci, i uzyskane wyniki prognoz — procent obserwacji poprawnie zaklasyfikowanych — zestawione są w tabeli 3.15.

Tabela 3.15. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu RBF, zbiór treningowy, punkty klasy AB, C, D (16 zmiennych wejściowych, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*rbf01a.nno*	80.60
2.	*rbf02a.nno*	76.87
3.	*rbf03a.nno*	63.43
4.	*rbf04a.nno*	63.43
5.	*rbf05a.nno*	63.43
6.	*rbf06a.nno*	63.43
7.	*rbf07a.nno*	28.36

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.15:

- *rbf01a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf02a.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf03a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf04a.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf05a.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf06a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.5; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf07a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Gaussian*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*).

Sieć typu RBF daje lepsze wyniki prognoz niż sieć MLP, prawie 80% punktów RMWP zostało poprawnie zaklasyfikowanych do obszaru o określonym użytkowaniu terenu. Sam proces uczenia się sieci trwa znacznie krócej niż w przypadku sieci typu MPL.

3.2.1.3. Sieć typu Bayesa

Ponieważ sieć Bayesa nie korzysta ze zbioru walidacyjnego, zbiór danych wejściowych podzielono na podzbiór treningowy i testowy w proporcjach 90% : 10% (150 : 17 obserwacji).

Pierwszy model został zbudowany przy domyślnych ustawieniach modułu sieci Bayesa, kolejne modyfikacje dotyczyły grup parametrów i sposobu wyboru najlepszego modelu. Konfiguracje poszczególnych modeli sieci i uzyskane wyniki prognoz — procent obserwacji poprawnie zaklasyfikowanych — zestawione są w tabeli 3.16.

Tabela 3.16. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu Bayesa, zbiór treningowy punkty klasy AB, C, D (16 zmiennych wejściowych, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*bayes01a.nno*	88.00
2.	*bayes02a.nno*	88.00
3.	*bayes03a.nno*	66.67
4.	*bayes04a.nno*	88.00
5.	*bayes05a.nno*	88.00
6.	*bayes06a.nno*	92.00
7.	*bayes07a.nno*	92.00
8.	*bayes07a.nno*	91.33

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.16:

- *bayes01a.nno* — struktura sieci: 16–4–3 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; wybór modelu: *Most Likely Model*;
- *bayes02a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Separate Weight and Biases*; wybór modelu: *Most Likely Model*;
- *bayes03a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Auto Relevance Detection*; wybór modelu: *Most Likely Model*;
- *bayes04a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Single Weight Group*; wybór modelu: *Most Likely Model*;
- *bayes05a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; wybór modelu: *Commitee Decision*;
- *bayes06a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 5; wybór modelu: *Most Likely Model*;
- *bayes07a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 10; wybór modelu: *Most Likely Model*;
- *bayes08a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 10; wybór modelu: *Commitee Decision*.

Sieć typu Bayesa poprawnie klasyfikuje ok. 90% punktów.

3.2.1.4. Wybór najlepszego modelu sieci

Porównując wyniki prognoz (procent punktów RMWP poprawnie zaklasyfikowanych do obszaru o określonym sposobie użytkowania terenu) uzyskanych za pomocą modeli MLP, RBF i Bayesa (tab. 3.14, 3.15 i 3.16), nalepszym modelem pozwalającym na klasyfikowanie punktu do obszaru o określonym sposobie zagospodarowania terenu na podstawie oznaczeń wskaźni-

ków fizyko-chemicznych wód (największy procent obserwacji poprawnie zaklasyfikowanych) okazał się model sieci Bayesa, którego wyniki zapisane są w pliku *bayes07a.nno*.

Do modelu wczytano testowy, roboczy plik danych (plik *run01a.sav*, w którym są te same wartości na których sieć się „uczyła") w celu sprawdzenia zdolności modelu do klasyfikacji.

Wyniki klasyfikacji dla zbioru testowego zostały zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS (*output1a.sav*).

Sieć poprawnie prognozowała 88.6% obserwacji — punktów monitoringowych — ze zbioru roboczego.

3.2.2. Prognozy dla punktów RMWP o klasie zagrożenia AB (wariant 2.)

Kolejne testy przeprowadzono dla zbioru *zbior02a.sav* utworzonego wg wariantu 2. (tab. 3.1).

Zbiór danych wejściowych podzielono na podzbiory: treningowy, testowy i walidacyjny. Podzbiór treningowy obejmuje 80% obserwacji (121 obserwacji), w podzbiorach walidacyjnym i testowym jest po 10% obserwacji (15 obserwacji).

Testowano różne modele sieci z grupy sieci nadzorowanych, konfiguracje tych modeli i uzyskane dla zbioru trenigowego wyniki prognoz wskaźników chemicznych wód zestawione są w tabelach 3.17, 3.18 oraz 3.19.

3.2.2.1. Sieć typu MLP

Tabela 3.17. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu MLP, zbiór treningowy, punkty klasy AB (16 zmiennych wejściowych, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*mlp08a.nno*	87.60
2.	*mlp09a.nno*	90.08
3.	*mlp10a.nno*	89.26
4.	*mlp11a.nno*	80.17
5.	*mlp12a.nno*	82.64
6.	*mlp13a.nno*	79.34
7.	*mlp14a.nno*	78.51

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.17:

- *mlp08a.nno* — struktura sieci: 16–4–3 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp09a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp10a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp11a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;

- *mlp12a.nno* — struktura sieci: 16–4–3; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp13a.nno* — struktura sieci: 16–4–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp14a.nno* — struktura sieci: 16–4–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*).

3.2.2.2. Sieć typu RBF

Zachowano podział zbioru danych wejściowych jak w przypadku sieci MLP: podzbiór treningowy (80% obserwacji — 121 obserwacji), walidacyjny i testowy (po 10% — 15 obserwacji).

Tabela 3.18. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu RBF, zbiór treningowy, punkty klasy AB (16 zmiennych wejściowych, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*rbf08a.nno*	81.82
2.	*rbf09a.nno*	92.56
3.	*rbf10a.nno*	66.12
4.	*rbf11a.nno*	66.12
5.	*rbf12a.nno*	66.12
6.	*rbf13a.nno*	74.38
7.	*rbf14a.nno*	29.75

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.18:

- *rbf08a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf09a.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf10a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf11a.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf12a.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf13a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.5; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf14a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Gaussian*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*).

3.2.2.3. Sieć typu Bayesa

Sieć Bayesa nie korzysta ze zbioru walidacyjnego, dokonano więc podziału zbioru danych wejściowych na podzbiór treningowy i testowy w proporcjach 90% : 10% (136 : 15 obserwacji).

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.19:

- *bayes08a.nno* — struktura sieci: 16–4–3 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; wybór modelu: *Most Likely Model*;

- *bayes09a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Separate Weight and Biases*; wybór modelu: *Most Likely Model*;
- *bayes10a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Auto Relevance Detection*; wybór modelu: *Most Likely Model*;
- *bayes11a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Single Weight Group*; wybór modelu: *Most Likely Model*;
- *bayes12a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; wybór modelu: *Committee Decision*;
- *bayes13a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 5; wybór modelu: *Most Likely Model*;
- *bayes14a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 10; wybór modelu: *Most Likely Model*;
- *bayes15a.nno* — struktura sieci: 16–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 10; wybór modelu: *Commitee Decision*.

Tabela 3.19. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu Bayesa, zbiór treningowy, punkty klasy AB (16 zmiennych wejściowych, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*bayes08a.nno*	91.91
2.	*bayes09a.nno*	91.18
3.	*bayes10a.nno*	91.91
4.	*bayes11a.nno*	91.91
5.	*bayes12a.nno*	91.91
6.	*bayes13a.nno*	94.12
7.	*bayes14a.nno*	97.06
8.	*bayes15a.nno*	97.06

3.2.2.4. Wybór najlepszego modelu sieci

Porównując wyniki prognoz (procent punktów RMWP klasy AB poprawnie zaklasyfikowanych do obszaru o określonym użytkowaniu terenu) uzyskanych za pomocą modeli MLP, RBF i Bayesa (tab. 3.17, tab. 3.18 i 3.19), nalepszym modelem pozwalającym na klasyfikowanie punktu do obszaru o określonym zagospodarowaniu na podstawie oznaczeń wskaźników fizyko-chemicznych wód (największy procent obserwacji poprawnie zakwalifikowanych) okazał się model sieci Bayesa, którego wyniki zapisane są w pliku *bayes14a.nno*.

Do modelu wczytano testowy, roboczy plik danych (plik *run02a.sav*, w którym są te same wartości na których sieć się „uczyła") w celu sprawdzenia zdolności modelu. Wyniki klasyfikacji dla zbioru testowego zostały zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS (*output2a.sav*).

Sieć poprawnie zakwalifikowała 91.4% obserwacji (punktów monitoringowych) ze zbioru roboczego, zatem po ograniczeniu obserwacji w zbiorze treningowym do punktów RMWP o klasie zagrożenia AB uzyskano znacznie lepsze wyniki prognoz.

3.2.3. Prognozy dla punktów RMWP o klasie zagrożenia AB z ograniczoną liczbą zmiennych (wariant 3.)

Kolejne testy przeprowadzono dla pliku (*zbior03a.sav*) z sześcioma zmiennymi typu wejściowego, utworzonym wg wariantu 3. (tab. 3.1).

Zbiór ten po wczytaniu do programu Neural Connection podzielono na podzbiory: treningowy — obejmujący 80% obserwacji (114 obserwacji), walidacyjny — obejmujący 10% obserwacji (15 obserwacji) i testowy — (14 obserwacji).

Przy siedmiu zmiennych typu wejściowego ($M = 6$) i jednej zmiennej docelowej ($N = 1$) w zbiorze powinno być co najmniej $10(M + N) = 10(6 + 1) = 70$ obserwacji (SPSS, 1997), zatem w analizowanym przypadku warunek ten jest spełniony.

Testowano różne modele sieci, podobnie jak w przypadku zbiorów tworzonych wg wariantów 1. i 2., konfiguracje poszczególnych modeli i uzyskane dla zbioru trenigowego wyniki prognoz wskaźników chemicznych wód zestawione są w tabelach 3.20, 3.21, 3.22.

3.2.3.1. Sieć typu MLP

Tabela 3.20. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu MLP, zbiór treningowy, punkty klasy AB (6 zmiennych typu wejściowego, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*mlp15a.nno*	88.29
2.	*mlp16a.nno*	85.59
3.	*mlp17a.nno*	77.48
4.	*mlp18a.nno*	68.47
5.	*mlp19a.nno*	81.98
6.	*mlp20a.nno*	78.38
7.	*mlp21a.nno*	78.38

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.20:

- *mlp15a.nno* — struktura sieci: 6–4–3 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp16a.nno* — struktura sieci: 6–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp17a.nno* — struktura sieci: 6–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp18a.nno* — struktura sieci: 6–4–3; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;
- *mlp19a.nno* — struktura sieci: 6–4–3; funkcja aktywacji neuronów: sigmoid (*sigmoid*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po pełnym przebiegu (*epoch*); rozkład wag neuronów: jednostajny (*uniform*); automatycznie generowana liczba neuronów w warstwie ukrytej;

- *mlp20a.nno* — struktura sieci: 6–4–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: najszybszego spadku (*steepest descent*); metoda uaktualniania wag: po każdym kolejnym etapie (*pattern*); rozkład wag neuronów: jednostajny (*uniform*);
- *mlp21a.nno* — struktura sieci: 6–4–4–3; funkcja aktywacji neuronów: tangens hiperboliczny (*tanh*); algorytm uczący: gradient sprzężony (*conj. gradient*); rozkład wag neuronów: jednostajny (*uniform*).

3.2.3.2. Sieć typu RBF

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.21:

- *rbf15a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf16a.nno* — odległość: *City Block*; funkcja nieliniowa: *Spline*; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf17a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf18a.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.1; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf19a.nno* — odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.2; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*);
- *rbf20a.nno* — odległość: *Euclidean*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0.5; liczba centrów: 5; rozkład centrów: oparty na danych (*Sample*).

Tabela 3.21. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu RBF, zbiór treningowy, punkty klasy AB (6 zmiennych typu wejściowego, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*rbf15a.nno*	74.77
2.	*rbf16a.nno*	72.97
3.	*rbf17a.nno*	69.37
4.	*rbf18a.nno*	72.07
5.	*rbf19a.nno*	67.57
6.	*rbf20a.nno*	75.68

3.2.3.3. Sieć typu Bayesa

Sieć Bayesa nie korzysta ze zbioru walidacyjnego, dokonano więc podziału zbioru danych wejściowych na podzbiór treningowy i testowy w proporcjach 90% : 10% (129 : 14 obserwacji).

Charakterystyka badanych struktur sieci zestawionych w tabeli 3.22:

- *bayes16a.nno* — struktura sieci: 6–4–3 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej); automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; wybór modelu: *Most Likely Model*;
- *bayes17a.nno* — struktura sieci: 6–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Separate Weight and Biases*; wybór modelu: *Most Likely Model*;
- *bayes18a.nno* — struktura sieci: 6–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Auto Relevance Detection*; wybór modelu: *Most Likely Model*;
- *bayes19a.nno* — struktura sieci: 6–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Single Weight Group*; wybór modelu: *Most Likely Model*;

- *bayes20a.nno* — struktura sieci: 6–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; wybór modelu: *Commitee Decision*;
- *bayes21a.nno* — struktura sieci: 6–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 5; wybór modelu: *Most Likely Model*;
- *bayes22a.nno* — struktura sieci: 6–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 10; wybór modelu: *Most Likely Model*;
- *bayes23a.nno* — struktura sieci: 6–4–3; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Weight Grouped by Layer*; liczba modeli: 10; wybór modelu: *Commitee Decision*.

Tabela 3.22. Procent poprawnie zaklasyfikowanych obserwacji — prognozy sposobu zagospodarowania terenu na podstawie oznaczeń wskaźników jakości wód — sieć typu Bayesa, zbiór treningowy, punkty klasy AB (6 zmiennych typu wejściowego, 1 zmienna docelowa)

Lp.	Nazwa zbioru wynikowego	Procent poprawnych prognoz [%]
1.	*bayes16a.nno*	81.60
2.	*bayes17a.nno*	80.00
3.	*bayes18a.nno*	36.80
4.	*bayes19a.nno*	81.60
5.	*bayes20a.nno*	81.60
6.	*bayes21a.nno*	80.80
7.	*bayes22a.nno*	80.80
8.	*bayes23a.nno*	80.00

3.2.3.4. Wybór najlepszego modelu sieci

Porównując wyniki prognoz uzyskanych za pomocą modeli MLP, RBF i Bayesa (tab. 3.20, tab. 3.21 i 3.22), dla punktów o klasie zagrożenia AB z ograniczoną liczbą zmiennych wejściowych, nalepszym modelem pozwalającym na klasyfikowanie punktu RMWP do obszaru o określonym użytkowaniu terenu w oparciu o oznaczenia wskaźników fizyko-chemicznych wód (największy procent obserwacji poprawnie zakwalifikowanych) okazał się model sieci MLP, którego wyniki zapisane są w pliku *mlp15a.nno*.

Do modelu wczytano testowy, roboczy plik danych (plik *run03a.sav*, w którym są te same wartości na których sieć się „uczyła") w celu sprawdzenia zdolności modelu. Wyniki klasyfikacji dla zbioru testowego zostały zapisane (za pomocą narzędzia *Data Output*) do pliku w formacie SPSS (*output3a.sav*).

Sieć poprawnie zakwalifikowała 84.9% obserwacji ze zbioru roboczego, zatem po ograniczeniu zmiennych typu wejściowego, na podstawie których sieć prognozuje sposób zagospodarowania terenu uzyskano gorsze wyniki — mniejszy procent punktów RMWP poprawnie zaklasyfikowanych.

Podsumowanie

Potrzeba optymalizacji gęstości opróbowania sieci monitoringowych jakości wód podziemnych wynika z jednej strony z ograniczonych środków finansowych na badania monitoringowe, a z drugiej — z konieczności zapewnienia takiego układu (rozmieszczenia) i gęstości punktów tworzących sieć monitoringową, która zapewni rejestrację zmian jakości wód w układzie przestrzennym w zlewni, jednostce hydrogeologicznej czy też zbiorniku wód podziemnych.

Rozważane w niniejszej pracy podejście wykorzystujące metody sieci neuronowych do prognozowania zmian jakości wód w układzie przestrzennym oparte zostało na istniejącej bazie danych, zawierającej wyniki uzyskane w ramach regionalnego monitoringu jakości wód podziemnych RMWP przeprowadzonego dla zlewni górnej Wisły w latach 1993–1994 (Witczak et al., 1994).

Sieć RMWP dorzecza górnej Wisły składa się ze 172 punktów RMWP, z czego w obszarze RZGW Kraków znajduje się 117, zaś w obszarze RZGW Katowice — 55 punktów RMWP. Analizie poddano wyniki badań jakości wód podziemnych pobranych w I serii opróbowania (okres mokry, V–IX 1993). W serii tej opróbowaniem i analizą objęto 167 punktów RMWP, gdyż punkty 11012, 21024, 21047, 21052 i 21060 (wg numeracji punktów w bazie MONBADA), ze względu na niezakończony proces ich adaptacji nie zostały opróbowane (Witczak et al., 1994).

Wyniki oznaczeń terenowych i laboratoryjnych 55 wskaźników fizyko-chemicznych wód poddano weryfikacji. Podstawę do tej weryfikacji stanowiły dane zgromadzone w trakcie terenowego programu kontroli jakości QA/QC prowadzonego równolegle z opróbowaniem sieci monitoringowej RMWP dorzecza górnej Wisły.

Dane poddano weryfikacji na trzy sposoby: wyznaczono granice oznaczalności badanych wskaźników (laboratoryjną DL i praktyczną PDL), oszacowano udział wariancji technicznej σ^2_{tech} w wariancji całkowitej σ^2_{tot} na podstawie wyników oznaczeń próbek dublowanych, z wykorzystaniem klasycznej analizy wariancji ANOVA oraz elastycznego postępowania statystycznego (*robust statistics*) oraz dokonano analizy rozkładu tych wskaźników.

Praktyczna granica oznaczalności PDL ma znaczenie szacunkowe, informuje od jakiego stężenia można oczekiwać, że w warunkach rutynowego opróbowania i w prawidłowo wyposażonym laboratorium, przy zastosowaniu określonej metody oznaczeń analitycznych, uzyska się zadowalającą precyzję wyników. PDL powinna mieć wartość jak najbliższą laboratoryjnej granicy oznaczalności DL; w idealnym przypadku PDL = DL (PDL/DL = 1).

W przypadku rtęci stosunek PDL/DL $\approx$ 18.5, co świadczy o niskiej precyzji oznaczeń tego wskaźnika. Podobna sytuacja ma miejsce w przypadku oznaczeń chloroformu (PDL/DL $\approx$ 39900). Z kolei dla makroskładników: sodu, chlorków i siarczanów stosunek PDL/DL = 1, co oznacza, że wyniki te cechują się zadowalającą precyzją.

Uzyskane wyniki potwierdziły konieczność prowadzenia kontroli wartości PDL, a w razie niezadowalających wyników, potrzebę wykrycia błędów grubych i ich usunięcia, tak by zapewnić właściwy poziom PDL.

Terenowy program kontroli jakości QA/QC prowadzony w zlewni górnej Wisły umożliwił też weryfikację wyników oznaczeń wskaźników fizyko-chemicznych wód poprzez ob-

liczenie tzw. wariancji technicznej (σ^2_{tech}), uwzględniającej łączny wpływ błędów opróbowania i analityki. Obliczenia wariancji przeprowadzono za pomocą programu komputerowego ROB 2, pozwalającego na oszacowanie wariancji metodą klasyczną i metodą *robust statistics*. Ta ostatnia polega na zastosowaniu tzw. elastycznego postępowania statystycznego bez konieczności skomplikowanego odrzucania błędów grubych.

Dla sześciu wskaźników spośród czterdziestu trzech oznaczanych w laboratorium (bor, chrom ogólny, kadm, substancje powierzchniowo-czynne, benzo-a-piren, suma 6WWA) nie można było wyznaczyć poziomu wariancji technicznej ze względu na niedostateczną ($N < 11$) liczbę par próbek normalnych i dublowanych. W dwudziestu jeden przypadkach poziom wariancji technicznej wyznaczonej z wykorzystaniem klasycznej analizy wariancji ANOVA jest zadowalający ($\sigma^2_{tech} < 20\%$). W siedmiu przypadkach (glin, miedź, rtęć, chloroform, DDT, DDE, metoksychlor) wariancja techniczna kształtowała się na poziomie ok. 30% wariancji całkowitej. W przypadku potasu, żelaza ogólnego i azotu organicznego uzyskano wyniki w przedziale 40–60% zmienności całkowitej. Wariancją techniczną powyżej 60% wariancji całkowitej charakteryzowały się wyniki oznaczeń manganu, ołowiu i fenoli lotnych. Najniższą precyzją — wariancja techniczna stanowi ponad 80% wariancji całkowitej — charakteryzowały się wyniki oznaczeń cynku, czterochloroetylenu i trójchloroetylenu.

W wyniku elastycznego postępowania statystycznego (*robust statistics*) w większości przypadków uzyskano niższą wariancję techniczną, co oznacza, że wyniki pomiarów badanych wskaźników obarczone są błędami grubymi. W trzydziestu dwóch przypadkach poziom wariancji technicznej nie przekraczał 20% wariancji całkowitej. Oznaczenia azotu azotynowego, niklu i ołowiu charakteryzowała wariancja techniczna w granicach od 20 do 40%. Jeszcze większą zmiennością charakteryzowały się wyniki oznaczeń fenoli i czterochloroetylenu — wariancja techniczna stanowi w tym przypadku 60–70% wariancji całkowitej.

Wskaźniki chemiczne wód, dla których wariancja techniczna obliczona metodą klasyczną przekraczała dopuszczalny poziom 20% wyłączono ze zbioru, na którym oparto prognozowanie zmian jakości wód za pomocą sieci neuronowych, gdyż błędy w bazie danych wejściowych skutkują powielaniem ich w prognozach dotyczących zmian jakości wód. Wyłączono też z analizy obserwacje anomalne, obarczone błędami grubymi, i oznaczenia tych wskaźników, w których ponad 20% stanowiły wyniki poniżej granicy oznaczalności $<$ DL.

Dane literaturowe wskazują, że o jakości uzyskanych wyników pomiarów badanych wskaźników fizyko-chemicznych wód decydują głównie: proces opróbowania (Nielsen, 1991), precyzja zastosowanej metody analitycznej i warunki, w jakich wykonywane są oznaczenia — powtarzalność i odtwarzalność pomiarów (Huber, 1997).

To skłoniło autorkę do podjęcia dodatkowych badań, których obiektem były wody podziemne występujące w wapieniach górnej jury w obszarze miasta Krakowa (Zdrój Królewski). Do szczegółowych rozważań wybrano wyniki oznaczeń cynku (przedstawiciela metali ciężkich), który z uwagi na łatwość migracji powszechnie występuje w wodach podziemnych (Macioszczyk, 1987). Sprawdzano jak zmienia się jakość uzyskiwanych wyników w zależności od zastosowanej metodyki opróbowania:

- sprzęt do opróbowania wielokrotnego użytku firmy Eijkelkamp;
- sprzęt do opróbowania jednokrotnego użytku firmy Millipore;

metodyki oznaczeń:

- AAS z granicą oznaczalności cynku DL = 0.01 mg/dm^3;
- ICP-AES z granicą oznaczalności cynku DL = 0.005 mg/dm^3;

oraz warunków wykonywania analiz (ta sama metodyka oznaczeń w warunkach powtarzalności i odtwarzalności).

Z przeprowadzonych analiz wynika, iż zastosowanie sprzętu jednokrotnego użytku w miejsce sprzętu wielokrotnego użytku powoduje wzrost precyzji oznaczeń. W przypadku gdy proces opróbowania wody jurajskiej prowadzony był przez jednego operatora, z wykorzystaniem sprzętu do filtracji i opróbowania jednorazowego użytku Millipore, współczynnik

zmienności osiąga mniejszą wartość $V \approx 20$–50% niż w przypadku gdy opróbowanie prowadzono z wykorzystaniem sprzętu wielokrotnego użytku Eijkelkamp ($V \approx 100$–190%).

Również zastosowanie w badaniach metody o niższej granicy oznaczalności (metoda ICP-AES zamiast AAS) wpływa istotnie na zwiększenie precyzji uzyskiwanych oznaczeń — prawie dwukrotne zmniejszenie współczynnika zmienności V z ok. 90% w przypadku metody AAS do ok. 40% po zastosowaniu metody ICP-AES.

Na precyzję oznaczeń mają również wpływ warunki, w jakich wykonywane są analizy. Z przeprowadzonych obliczeń wynika, że precyzja oznaczeń w warunkach powtarzalności jest co najmniej 2–3-krotnie wyższa ($V \approx 20$–50%) niż precyzja wyznaczana w warunkach odtwarzalności ($V \approx 100\%$).

Aby zatem uzyskiwać wiarygodne wyniki oznaczeń wskaźników jakości wód podziemnych należy tak planować badania monitoringowe, by próbki wody pobierane były przy zastosowaniu **sprzętu jednokrotnego użytku**, filtrowane *on-line* w terenie, a następnie analizowane **w jednym laboratorium**, **w warunkach powtarzalności pomiarów**, przy zastosowaniu **metody pomiarowej o odpowiednio niskiej granicy oznaczalności** DL, 1–2 rzędy niższej w stosunku do spodziewanych stężeń w próbkach pobranych ze zbiornika wód podziemnych.

Z wykorzystaniem procedury eksploracji dostępnej w programie SPSS PL for Windows przeprowadzono analizę rozkładu 55 wskaźników fizyko-chemicznych oznaczanych w próbkach wody podziemnej pobranej z sieci RMWP dorzecza górnej Wisły, zidentyfikowano wartości ekstremalne (na wykresach typu „skrzynka z wąsami" i „łodyga i liście"). Na tej podstawie dokonano podziału badanego zbioru analiz na podzbiory, charakteryzujące subpopulacje: anomalną (obserwacje ekstremalne) i typową (obserwacje typowe, pozostałe w zbiorze po wyłączeniu z analizy obserwacji ekstremalnych).

Następnie ponownie wykonano analizę rozkładu badanych zmiennych dla subpopulacji typowej, i w niektórych przypadkach — gdy liczba próbek wyłączonych z analizy była większa od siedmiu, analizę opisową subpopulacji anomalnej.

W efekcie, w oparciu o obliczone parametry kontroli jakości (DL, PDL, σ^2_{tech}) i procedurę eksploracji zmiennych, w zweryfikowanej bazie danych, na podstawie której prowadzono próby prognozowania jakości wód podziemnych w układzie przestrzennym pozostało szesnaście zmiennych:

- temperatura [°C];
- odczyn pH;
- suma substancji rozpuszczonych [mg/dm^3];
- zasadowość ogólna [mval/dm^3];
- twardość ogólna [mg $CaCO_3$/dm^3];
- sód [mg/dm^3];
- magnez [mg/dm^3];
- wapń [mg/dm^3];
- chlorki [mg/dm^3];
- siarczany [mg/dm^3];
- krzemionka zdysocjowana [mg/dm^3];
- fluorki [mg/dm^3];
- cynk [mg/dm^3];
- współczynnik absorpcji UV (A 254);
- rozpuszczony węgiel organiczny [mg/dm^3];
- utlenialność ChZT-Mn [mg/dm^3].

Dla danych zweryfikowanych wykonano za pomocą programu GEO-EAS v. 1.2.1 ocenę geostatystyczną (metodą krigingu) i wyznaczono regionalne tło hydrogeochemiczne dla obszaru dorzecza górnej Wisły. Uzyskane wartości regionalnego tła hydrogeochemicznego obszaru zlewni górnej Wisły pokrywają się z wartościami ogólnopolskimi. Jedynie w przy-

padku siarczanów uzyskano wartości wyższe od ogólnopolskiego tła hydrogeochemicznego, co może świadczyć o niekorzystnym wpływie aglomeracji przemysłowej na jakość wód podziemnych omawianego obszaru.

Na zweryfikowanej bazie danych (w skład której wchodziło 16 wskaźników fizyko-chemicznych jakości wód) przeprowadzono próby predykcji wskaźników fizyko-chemicznych wód dla punktu monitoringowego o określonych współrzędnych oraz klasyfikacji punktu monitoringowego (na podstawie wyników oznaczeń wskaźników fizyko-chemicznych) do obszaru o określonym użytkowaniu terenu.

Próby te prowadzono dla trzech wariantów danych zweryfikowanych:

- zbiór zawierający wszystkie zweryfikowane wskaźniki fizyko-chemiczne (16) i punkty monitoringowe o klasach zagrożenia wód AB, C, D (167 punktów RMWP);
- zbiór zawierający wszystkie zweryfikowane wskaźniki fizyko-chemiczne (16), ale punkty monitoringowe ograniczone do klasy zagrożenia AB (151 punktów RMWP);
- zbiór zawierający punkty monitoringowe o klasie zagrożenia AB i 6 wskaźników zweryfikowanych (są to wskaźniki, w których wystąpiła najmniejsza liczba braków danych, $n \leqslant 5$).

Różne warianty danych wejściowych umożliwiły ocenę wpływu liczby zweryfikowanych wskaźników fizyko-chemicznych wód oraz liczby punktów monitoringowych (o różnym stopniu zagrożenia) w bazie danych na jakość uzyskiwanych prognoz.

Do rozwiązania zagadnień predykcji jakości wód podziemnych w danym nieopróbowanym punkcie monitoringowym na podstawie oznaczeń wskaźników fizyko-chemicznych w sąsiednich punktach opróbowanych wykorzystano modele sieci neuronowych z grupy sieci nadzorowanych (*supervised*). W programie Neural Connection z grupy tej dostępne są sieci: wielowarstwowy perceptron MLP, radialna funkcja bazowa RBF i sieć Bayesa. Jakość uzyskiwanych prognoz oceniano na podstawie średniego błędu względnego prognoz MA% (obliczanego wg wzoru (3.3)) dla zbiorów treningowych.

Następnie do modelu o najmniejszym względnym błędzie prognoz MA%, w tym przypadku we wszystkich wariantach był to model sieci RBF, wprowadzano dane zewnętrzne, tzw. robocze, by sprawdzić jakość uzyskiwanych z modelu prognoz w przypadku nowych danych wejściowych. Średni błąd względny uzyskanych prognoz kształtował się na poziomie od setnych części procenta do kilkunastu procent. Błąd ten był uzależniony od doboru zmiennych typu wejściowego. Największy błąd zaobserwowano dla pliku z 16 prognozowanymi zmiennymi i punktami RWMP o różnej klasie zagrożenia wód: AB, C, D (wariant 1.).

Po ograniczeniu obserwacji w zbiorze wejściowym do punktów RMWP o klasie zagrożenia AB (wariant 2. i 3.) zaobserwowano zmniejszenie wartości średniego błędu względnego prognoz. Współczynniki korelacji wartości obserwowanych z prognozowanymi kształtowały się na poziomie 0.383–0.904.

Nie stwierdzono związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników. Poziom tych błędów zależy jedynie od konfiguracji sieci.

Następnie na modelu RBF, dającym najlepsze wyniki prognoz, najmniejsze błędy względne prognoz MA% (6 zmiennych docelowych dla punktów RMWP klasy AB — wariant 3.), prowadzono próby optymalizacji gęstości opróbowania sieci RMWP, poprzez wyłączenie z istniejącej sieci pomiarowej części punktów i sprawdzenie czy na na podstawie pozostałych punktów — opróbowanych — można prognozować zmiany jakości wód podziemnych w całej sieci monitoringowej

W obszarze dorzecza górnej Wisły znajduje się 172 punkty RMWP, oznacza to, że 1 punkt RMWP przypada na 280.6 km^2, przy czym w obszarze RZGW Katowice jest to 1 pkt RMWP na 134.2 km^2, a w obszarze RZGW Kraków 1 pkt RMWP na 349.5 km^2.

W I serii opróbowania RMWP dorzecza górnej Wisły pobrano 167 próbek, co oznacza, że opróbowano średnio 1 punkt RMWP na 289 km^2. W analizowanym przypadku sieć monito-

ringowa została dodatkowo na wstępie sztucznie „rozgęszczona", przez wyłączenie z analizy punktów RMWP, w których wystąpiły braki danych (zatem gęstość sieci wynosi 1 punkt RMWP/337.6 km^2).

Próby prognozowania jakości wód w zlewni górnej Wisły prowadzono dla sieci o gęstości 1 pkt RMWP/419.7 km^2 — powstałej poprzez wyłączenie co piątego punktu z arkusza z danymi. Średni błąd względny prognoz kształtował się w tym przypadku na poziomie kilku–kilkudziesięciu procent, współczynnik korelacji wartości obserwowanych i prognozowanych zmieniał się od 0.204 do 0.785.

Następnie z sieci wyłączono te punkty, w pobliżu których (w odległości 10–15 km) znajdowały się co najmniej dwa inne punkty RMWP. Sieć monitoringowa miała wówczas gęstość 1 pkt RMWP/398.9 km^2. Dla takiego wariantu opróbowania średni błąd względny prognoz wskaźników fizyko-chemicznych we wszystkich przypadkach (za wyjątkiem krzemionki) kształtował się na poziomie kilku procent, współczynnik korelacji wartości obserwowanych i prognozowanych mieścił się w zakresie 0.569–0.860.

Oznacza to, że przy opróbowaniu np. 84% punktów monitoringowych w istniejącej sieci RMWP można z wykorzystaniem sieci neuronowych, na podstawie zweryfikowanych wyników oznaczeń wskaźników fizyko-chemicznych w tych punktach uzyskać wiarygodne prognozy dotyczące jakości wód w punktach nieopróbowanych.

Trudno wyznaczyć tu graniczną wartość, do jakiej można ograniczać liczbę opróbowanych, bądź nieopróbowanych punktów, czy też podać minimalną gęstość opróbowania sieci. Należy przeprowadzić co najmniej jedno pełne opróbowanie sieci o największej możliwej gęstości, wykonać pełne analizy wszystkich pobranych próbek, dokonać weryfikacji uzyskanych wyników oznaczeń i zbudować na tej podstawie model sieci neuronowej. Następnie należy na tym modelu eksperymentować, sprawdzając do jakiego momentu będzie się uzyskiwać wiarygodne prognozy jakości wód w badanym obszarze.

Sieci neuronowe wykorzystano też do rozwiązania zagadnień klasyfikacji. Sprawdzano, czy na podstawie wyników oznaczeń wskaźników jakości wód można uzyskać dane dotyczące sposobu użytkowania terenu w danym punkcie. Podobnie jak w przypadku zagadnień predykcji budowano różne modele sieci neuronowej z grupy sieci nadzorowanych (MLP, RBF, Bayesa), i sprawdzano, w przypadku której sieci największy procent punktów monitoringowych zostanie poprawnie zaklasyfikowany do obszaru o określonym użytkowaniu terenu.

Następnie do „najlepszego" modelu sieci (sieć typu Bayesa, najwięcej punktów RMWP poprawnie zaklasyfikowanych) wczytano testowy, roboczy plik danych w celu sprawdzenia zdolności modelu. W zależności od przyjętego wariantu danych sieć poprawnie prognozowała od 84.9–91.4% obserwacji — punktów monitoringowych — ze zbioru roboczego.

Uzyskane wyniki badań wskazują, że nowe narzędzie, jakim są **sieci neuronowe, można z powodzeniem wykorzystać do prognozowania zmian jakości wód w układzie przestrzennym**. Warunkiem jednak, by uzyskiwane prognozy były wiarygodne, jest konieczność **weryfikacji danych wejściowych wprowadzanych do modelu**.

Weryfikację baz danych zawierających oznaczenia wskaźników fizyko-chemicznych wód można prowadzić poprzez **analizę parametrów kontroli jakości uzyskiwanych oznaczeń** (DL, PDL, σ^2_{tech}) oraz **analizę ich rozkładu** (statystyki opisowe, percentyle, wykresy typu „łodyga i liście", „skrzynka z wąsami", histogram rozkładu, test i wykres normalności rozkładu, karta kontrolna pojedynczych pomiarów).

Zweryfikowane bazy danych hydrogeochemicznych mogą być podstawą do tworzenia modeli opisujących zmienność przestrzenną jakości wód podziemnych. Na ich podstawie można też prowadzić optymalizację gęstości opróbowania sieci monitoringowych, co przyczyni się do **obniżenia kosztów badań monitoringowych** oraz **skrócenia czasu ich prowadzenia**, **bez utraty informacji o zmianach jakości wód podziemnych** w układzie przestrzennym i czasowym (podstawa racjonalnego wykorzystania wód).

dodatek A

Charakterystyka punktów regionalnego monitoringu jakości wód podziemnych dorzecza górnej Wisły

Tabela A.1. Charakterystyka punktów regionalnego monitoringu jakości wód podziemnych (RMWP) dorzecza górnej Wisły w obszarze RZGW Katowice (wg Witczak et al., 1994)

Lp.	Numer punktu RMWP	Numer identyfikacyjny punktu w bazie MONBADA (vide rys. A.1)	Numer GZWP	Podobszar	Miejscowość	Województwo	Stratygrafia	Użytkowanie terenu	Klasa zagrożenia wód podziemnych
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1.	1	11001	327	KA	Tarnowskie Góry-Zyglin	KA	T2	L	AB
2.	2	11002	327	KA	Mierzęcice Łubne	KA	T2	R/L	AB
3.	3	11003	454	KA	Podwarpie	KA	T2	R	AB
4.	4	11004	454	KA	Czekanka	KA	T2	L/R	C
5.	5	11005	454	KA	Ciągowice	KA	T2	R	D
6.	6	11006	327	KA	Rogoźnik	KA	T2	L	AB
7.	7	11007	454	KA	Dąbrowa Górnicza-Ujejście	KA	T2	L/R	AB
8.	8	11008	452	KA	Niegowonice	KA	T2	R	D
9.	9	11009	454	KA	Hutki-Kanki	KA	T2	R	D
10.	10	11010	326	KA	Kwaśniów Górny	KA	J3	R	AB
11.	11	11011	326	KA	Dłużec	KA	J3	R	AB
12.	12	11012	326	KA	Zasępiec	KA	J3	R	AB
13.	13	11013	329	KA	Bytom	KA	T2	O-P	AB
14.	14	11014	329	KA	Bytom	KA	T2	O-P	AB
15.	15	11015	330	KA	Czekanów-Szałsza	KA	T2	R	AB
16.	16	11016	330	KA	Gliwice	KA	T2	L	AB
17.	17	11017	330	KA	Gliwice	KA	T	O-P	AB
18.	18	11018	331	KA	Ruda Śląska	KA	Q	O-P	AB
19.	19	11019	329	KA	Będzin-Małobądz	KA	T	R	AB
20.	20	11020	454	KA	Sławków	KA	T2	R	AB
21.	21	11021	326	KA	Braciejówka	KA	J3	R	AB
22.	22	11022	–	KA	Mysłowice-Brzezinka	KA	C	R	AB
23.	23	11023	452	KA	Mysłowice-Dziećkowice	KA	T2	R	AB
24.	24	11024	452	KA	Jaworzno	KA	T	R	AB
25.	25	11025	452	KA	Jaworzno	KA	T	R	AB
26.	26	11026	453	KA	Jaworzno-Szczakowa	KA	Q	O-P	AB
27.	27	11027	457	KA	Trzebinia	KA	C	L/R	AB
28.	28	11028	454	KA	Lgota	KA	T	R	AB
29.	29	11029	454	KA	Olkusz	KA	T2	O-P	AB
30.	30	11030	326	KA	Zederman	KA	J3	R	AB
31.	31	11031	452	KA	Chrzanów-Borowiec	KA	T	R	AB
32.	32	11032	449	KA	Babice-Wygiełzów	KA	Q	R	AB
33.	33	11033	349	KA	Jastrzębie	KA	Q	R	C
34.	34	11034	346	KA	Czarków	KA	Q	R	AB
35.	35	11035	346	KA	Pszczyna	KA	Q	R/O-P	C
36.	36	11036	449	KA	Zaborze	BB	Q	R	C
37.	37	11037	449	KA	Przeciszów	BB	Q	R	AB
38.	38	11038	347	KA	Zaborze-Gołysz	BB	Q	R	AB
39.	39	11039	447	KA	Inwałd	BB	K	L	AB
40.	40	11040	–	KA	Jeleśnia-Janiki	BB	X	L	AB
41.	41	11041	–	KA	Ogrodzona	BB	K	R	AB
42.	42	11042	348	KA	Ustroń na Podlesiu	BB	K	L	AB
43.	43	11043	348	KA	Brenna-Leśnica	BB	K	R	AB
44.	44	11044	348	KA	Bystra	BB	K	L	AB
45.	45	11045	447	KA	Bielsko-Biała	BB	K	L	AB
46.	46	11046	348	KA	Lipowa	BB	K	L	AB
47.	47	11047	447	KA	Targanice Kubasówka	BB	K	L	AB
48.	48	11048	445	KA	Korbielów	BB	X	L	AB

Tabela A.1 cd.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
49.	49	11049	445	KA	Grzechynia	BB	X	L	AB
50.	50	11050	445	KA	Osielec	BB	X	R	AB
51.	51	11051	445	KA	Sidzina, Rola Kamycka	BB	X	L	AB
52.	52	11052	445	KA	Zawoja, Przełęcz Krowiarki	BB	X	L	AB
53.	53	11053	348	KA	Kamesznica-Złatna	BB	KX	L	AB
54.	54	11054	445	KA	Żabnica	BB	X	L	AB
55.	55	11055	445	KA	Rajcza Dolna	BB	X	L	AB

Tabela A.2. Charakterystyka punktów regionalnego monitoringu jakości wód podziemnych (RMWP) dorzecza górnej Wisły w obszarze RZGW Kraków (wg Witczak et al., 1994)

Lp.	Numer punktu RMWP	Numer identyfikacyjny punktu w bazie MONBADA (vide rys. A.1)	Numer GZWP	Podobszar	Miejscowość	Województwo	Stratygrafia	Użytkowanie terenu	Klasa zagrożenia wód podziemnych
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1.	1	21001	409	KR-W	Wierzchowisko	KA	J3	R	AB
2.	2	21002	409	KR-W	Ulina Mała	KR	K2	R	AB
3.	3	21003	326	KR-W	Zadroże	KR	J3	R	AB
4.	4	21004	409	KR-W	Gołcza	KR	J3	R	AB
5.	5	21005	409	KR-N	Biskupice	KI	K2	R	AB
6.	6	21006	409	KR-W	Celiny	KR	K2	R	AB
7.	7	21007	409	KR-W	Maszków	KR	K2	R	AB
8.	8	21008	409	KR-W	Bibice	KR	K2	R	C
9.	9	21009	451	KR-W	Rajsko	KR	X	R	C
10.	10	21010	443	KR-W	Myślenice	KR	Q	OP	AB
11.	11	21011	445	KR-W	Skomielna Biała	NS	X	L	AB
12.	12	21012	439	KR-W	Orawka	NS	X	R	C
13.	13	21013	440	KR-W	Jabłonka	NS	Q	R	AB
14.	15	21015	414	KR-N	Ruda Strawczyńska	KI	T2	R	AB
15.	16	21016	414	KR-N	Strawczyn	KI	T1	R	AB
16.	17	21017	414	KR-N	Ćmińsk	KI	T1	R	AB
17.	19	21019	414	KR-N	Zagnańsk	KI	D2	R	AB
18.	20	21020	414	KR-N	Zagnańsk (Zachełmie)	KI	T1	R	AB
19.	21	21021	418	KR-N	Miedzianka	KI	D2	R	AB
20.	22	21022	418	KR-N	Nowiny	KI	D2	LP	AB
21.	23	21023	417	KR-N	Jaworznia	KI	D2	R	AB
22.	24	21024	418	KR-N	Czerwona Góra	KI	D2	RL	AB
23.	25	21025	417	KR-N	Zalesie	KI	D2	R	AB
24.	27	21027	417	KR-N	Białogon	KI	D2	R(OP)	AB
25.	28	21028	417	KR-N	Kielce	KI	D2	OP	AB
26.	29	21029	418	KR-N	Trzuskawica	KI	D2	R	AB
27.	30	21030	418	KR-N	Dyminy	KI	D2	RP	AB
28.	31	21031	418	KR-N	Marzysz	KI	D2	R	AB
29.	32	21032	418	KR-N	Borków	KI	D2	L	AB
30.	33	21033	416	KR-N	Bocheniec	KI	J3	L	AB
31.	34	21034	416	KR-N	Tokarnia	KI	J3	R	AB
32.	35	21035	409	KR-N	Kanice	KI	K2	R	AB
33.	36	21036	409	KR-N	Jędrzejów Piaski	KI	K2	R	AB
34.	37	21037	409	KR-N	Węgleniec	KI	K2	R	AB
35.	38	21038	409	KR-N	Brzeście	KI	K2	R	AB
36.	39	21039	409	KR-N	Szarbków	KI	K2	R	AB

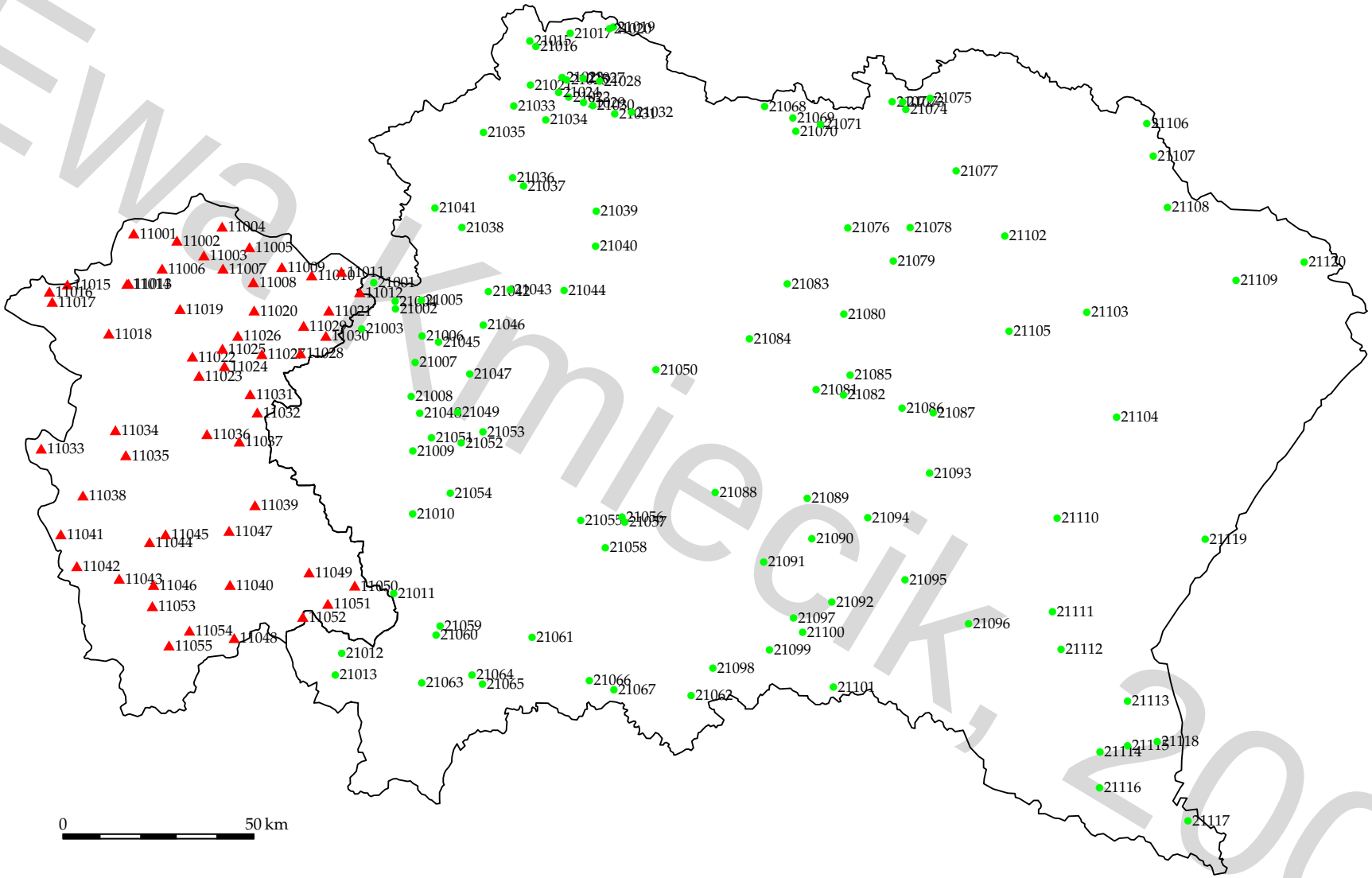
Tabela A.2 cd.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
37.	40	21040	409	KR-N	Marzęcin	KI	K2	R	AB
38.	41	21041	409	KR-N	Sędziszów	KI	K2	OP	AB
39.	42	21042	409	KR-N	Kropidło	KI	K2	R	AB
40.	43	21043	409	KR-N	Płużki	KI	K2	R	AB
41.	44	21044	409	KR-N	Mękarzewice	KI	K2	R	AB
42.	45	21045	409	KR-W	Słomniki	KR	K2	R	AB
43.	46	21046	409	KR-W	Smoniowice	KR	K2	R	AB
44.	47	21047	409	KR-W	Polekarcice	KR	K2	R	AB
45.	48	21048	450	KR-W	Kraków – Nowa Huta	KR	Q	OP	AB
46.	49	21049	450	KR-W	Kraków – Nowa Huta	KR	Q	OP	AB
47.	50	21050	–	KR-E	Gorzyce	TR	Q	R	AB
48.	51	21051	451	KR-W	Bieżanów	KR	X	OP	C
49.	52	21052	451	KR-W	Podłęże	KR	X	R	D
50.	53	21053	451	KR-W	Niepołomice	KR	X	R	D
51.	54	21054	443	KR-W	Dziekanowice	KR	Q	R	AB
52.	55	21055	–	KR-E	Iwkowa	TR	K	L	AB
53.	56	21056	435	KR-E	Filipowice	TR	Q	R	AB
54.	57	21057	436	KR-E	Filipowice	TR	K	LR	AB
55.	58	21058	436	KR-E	Rożnów	NS	K	L	AB
56.	59	21059	439	KR-W	Poręba Wielka-Koninki	NS	X	L	AB
57.	60	21060	439	KR-W	Poręba Wielka-Tobołów	NS	X	L	AB
58.	61	21061	437	KR-E	Czerniec	NS	X	R	AB
59.	62	21062	–	KR-E	Krynica	NS	X	L	AB
60.	63	21063	440	KR-E	Nowy Targ	NS	Q	RL	AB
61.	64	21064	440	KR-E	Dębno	NS	Q	R	AB
62.	65	21065	440	KR-E	Frydman	NS	Q	R	AB
63.	66	21066	438	KR-E	Rytro	NS	X	L	AB
64.	67	21067	438	KR-E	Piwniczna	NS	X	R	AB
65.	68	21068	421	KR-N	Boćkowice	TB	D2	R	AB
66.	69	21069	421	KR-N	Kobylany	TB	D2	R	AB
67.	70	21070	421	KR-N	Mydłowiec	TB	D2	R	AB
68.	71	21071	421	KR-N	Włostów	TB	D2	R	AB
69.	72	21072	422	KR-N	Pisary	TB	J3	R	AB
70.	73	21073	422	KR-N	Wygoda	TB	J3	R	AB
71.	74	21074	422	KR-N	Romanówka	TB	J2	R	AB
72.	75	21075	422	KR-N	Zawichost	TB	–	R(OP)	AB
73.	76	21076	–	KR-E	Baranów Sandomierski	TB	Q	R	AB
74.	77	21077	425	KR-E	Zbydniów	TB	Q	R	AB
75.	78	21078	425	KR-E	Studzieniec	TB	Q	R	AB
76.	79	21079	425	KR-E	Nowa Dęba	TB	Q	L	AB
77.	80	21080	425	KR-E	Szydłowiec	RZ	Q	L	AB
78.	81	21081	–	KR-E	Nagoszyn	TR	Q	R	AB
79.	82	21082	425	KR-E	Pustków	TR	Q	LP	AB
80.	83	21083	424	KR-E	Borowa	RZ	Q	R	AB
81.	84	21084	–	KR-E	Jamy	TR	Q	RL	AB
82.	85	21085	425	KR-E	Biały Bór	RZ	Q	LR	AB
83.	86	21086	–	KR-E	Krzywa	RZ	Q	R	AB
84.	87	21087	425	KR-E	Trzciana	RZ	Q	R	AB
85.	88	21088	434	KR-E	Tuchów	TR	Q	R	AB
86.	89	21089	433	KR-E	Brzostek	TR	Q	R/OP	AB
87.	90	21090	433	KR-E	Kołaczyce	KŚ	Q	R	AB
88.	91	21091	–	KR-E	Biecz	KŚ	X	R	C
89.	92	21092	433	KR-E	Osiek Jasielski	KŚ	QX	R	AB
90.	93	21093	432	KR-E	Czudec	RZ	Q	OP	AB
91.	94	21094	432	KR-E	Frysztak	RZ	Q	R	AB
92.	95	21095	432	KR-E	Krosno	KŚ	QX	OP	AB
93.	96	21096	432	KR-E	Besko	KŚ	QX	R	AB
94.	97	21097	–	KR-E	Folusz	KŚ	X	L	AB
95.	98	21098	–	KR-E	Śnietnica	NS	X	L	AB
96.	99	21099	–	KR-E	Gładyszów	NS	X	L	AB
97.	100	21100	–	KR-E	Folusz	KŚ	X	L	AB
98.	101	21101	–	KR-E	Ciechania	KŚ	X	L	AB

Tabela A.2 cd.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
99.	102	21102	425	KR-E	Barce	TB	Q	L	AB
100.	103	21103	425	KR-E	Przychojec	RZ	Q	R	AB
101.	104	21104	–	KR-E	Przeworsk	PR	Q	P-O/R	C
102.	105	21105	427	KR-E	Turza	RZ	Q	LR	AB
103.	106	21106	–	KR-E	Zagrody	ZA	K2	R	AB
104.	107	21107	–	KR-E	Frampol	ZA	K	R	C
105.	108	21108	428	KR-E	Biłgoraj	ZA	Q	R	C
106.	109	21109	428	KR-E	Łukowa	ZA	Q	R	AB
107.	110	21110	430	KR-E	Bachórz	PR	Q	R	AB
108.	111	21111	430	KR-E	Tyrawa Solna	KŚ	Q	RL	AB
109.	112	21112	430	KR-E	Załuż	KŚ	Q	R	AB
110.	113	21113	431	KR-E	Łobzew Górny	KŚ	X	R	AB
111.	114	21114	431	KR-E	Terka	KŚ	X	R	AB
112.	115	21115	431	KR-E	Olchowiec	KŚ	X	RL	AB
113.	116	21116	431	KR-E	Kalnica	KŚ	X	L	AB
114.	117	21117	431	KR-E	Muczne	KŚ	X	L	AB
115.	118	21118	431	KR-E	Czarna Polana	KŚ	X	L	AB
116.	119	21119	–	KR-E	Przemyśl	PR	Q	PO	AB
117.	120	21120	–	KR-E	Susiec	ZA	K	RL	AB

Objaśnienia skrótów i symboli: województwo: BB — bielsko-bialskie; KA — katowickie; KI — kieleckie; KR — krakowskie; KŚ — krośnieńskie; NS — nowosądeckie; PR — przemyskie; RZ — rzeszowskie; TB — tarnobrzeskie; TR — tarnowskie; ZA — zamojskie; **stratygrafia:** Q — czwartorzęd; X — trzeciorzęd; K — kreda; J — jura; T — trias; P — perm; C — karbon; D — dewon; S — sylur; O — ordowik; Y — kambr; N — kenozoik; M — mezozoik; F — paleozoik; E — prekambr; **numer GZWP** — numer Głównego Zbiornika Wód Podziemnych; **podobszar:** KA — obszar RZGW Katowice; KR — obszar RZGW Kraków (podobszary: KR-E — wschodni; KR-W — zachodni; KR-N — północny); **użytkowanie terenu:** R — rolnicze; L — leśne; O-P — osiedlowo-przemysłowe; **klasa zagrożenia:** AB — (czas migracji do 25 lat) wody zagrożone; C — (czas migracji od 25 do 100 lat) wody słabo zagrożone; D — (czas migracji ponad 100 lat) wody praktycznie niezagrożone.



Rysunek A.1. Regionalny monitoring jakości wód podziemnych (RMWP) w zlewni górnej Wisły, lokalizacja punktów monitoringowych (▲ — RZGW Katowice, ● — RZGW Kraków). Numeracja punktów zgodna z ich identyfikacją w bazie MONBADA

dodatek B

Ważniejsze pojęcia i definicje

Algorytm uczący (*learning algorithm*) — reguła określająca w jaki sposób ma przebiegać proces uczenia sieci neuronowej. Rodzaj wybranego algorytmu ma wpływ na to, jak dobrze sieć neuronowa rozwiązuje dany problem.

Analiza podstawowa (*base analysis*) — analiza stosowana przy systematycznych, powtarzalnych badaniach w sieciach monitoringu krajowego i regionalnego oraz jako pierwsze badanie w sieciach monitoringu lokalnego. Obejmuje oznaczenia azotu amonowego, azotanów, azotynów, barwy, chlorków, elektrolitycznej przewodności właściwej, fluorków, magnezu, manganu, odczynu pH, potasu, siarczanów, sodu, suchej pozostałości, twardości ogólnej, wapnia, zasadowości i żelaza.

Analiza szczegółowa (*detail analysis*) — stosowana w sieciach monitoringu krajowego i regionalnego jeden raz w roku, w uzasadnionych przypadkach także w sieciach monitoringu lokalnego (również 1 raz w roku). Obejmuje oznaczenia wskaźników włączonych do analizy podstawowej oraz co najmniej 50% wskaźników spośród innych będących podstawą klasyfikacji wód.

Analiza śladowa (*trace analysis*) — wykrywanie i/lub oznaczanie składników występujących w próbce w stężeniach poniżej 100 ppm.

Analiza wariancji (*analysis of variance*) — ogólne zróżnicowanie zbioru obserwacji mierzone sumą kwadratów odchyleń względem wartości średnich; może być ono podzielone na składniki odpowiadające różnym źródłom zróżnicowania przyjętym jako kryteria klasyfikacyjne zbioru obserwacji.

Analiza wody (*water analysis*) — oznacza zarówno czynność jak i wynik przeprowadzonych badań — określania składu chemicznego, cech fizycznych, organoleptycznych i bakteriologicznych wody. Podczas analizowania wody bada się substancje występujące w wodzie. Pojęcie analizy wody rozszerzone przymiotnikowo może określać zakres badań (np. analiza fizyko-chemiczna wody, analiza bakteriologiczna), szczegółowość badań (analiza ilościowa, analiza wskaźnikowa), miejsce przeprowadzonych badań (analiza terenowa) lub zastosowaną metodę badań.

Analiza wskaźnikowa (*traser analysis*) — wykonuje się ją w sieciach monitoringu lokalnego wokół ognisk zanieczyszczeń a w sieciach monitoringu lokalnego wokół ujęć wód podziemnych. Analiza ta obejmuje oznaczenia wybranego zespołu wskaźników. Ich liczbę i zestaw należy dostosować do lokalnych warunków hydrogeochemicznych i potrzeb badawczych.

Badania (*study*) — działanie techniczne, które polega na określeniu jednej lub wielu cech danego wyrobu, procesu lub usługi zgodnie z ustaloną procedurą.

Badania hydrogeochemiczne (*hydrogeochemical investigations*) — ich podstawowym celem jest określenie jakości wód, form występowania składników, zasad ich klasyfikacji, stopnia degradacji i ochrony jakości.

Badania hydrogeologiczne (*hydrogeological investigations*) — ich podstawowym celem jest określanie ilości i jakości wód użytkowych, zasad ich eksploatacji i ochrony przed zanieczyszczeniami.

Błąd bezwzględny $\Delta(a)$ (*absolute error*) — bezwzględne odchylenie pewnej obserwacji od jej „prawdziwej" wartości; wartość bezwzględna różnicy pomiędzy liczbą przybliżoną a i liczbą dokładną A:

$$\Delta(a) = |a - A|.$$

Błąd losowy (*random error*) — błąd, tzn. odchylenie od prawdziwej wartości, który ma własności zmiennej losowej w tym sensie, że każda konkretna wartość błędu występuje z pewnym prawdopodobieństwem określonym przez rozkład prawdopodobieństwa błędów:

$$x_i = \mu_x + e_i,$$

gdzie: x_i — i-ty pomiar określonej cechy; μ_x — wartość „prawdziwa" badanej cechy; e_i — błąd losowy i-tego pomiaru.

Błąd pierwszego rodzaju (*error of 1-st kind*) — α, błąd polegający na odrzuceniu hipotezy statystycznej, wówczas gdy powinna ona być przyjęta; uznaniu za niezgodną partii faktycznie zgodnej z wymaganiami lub procesu uregulowanego za nieuregulowany.

Błąd drugiego rodzaju (*error of 2-nd kind*) — β, błąd polegający na przyjęciu fałszywej hipotezy statystycznej; uznaniu za zgodną partii faktycznie niezgodnej z wymaganiami lub procesu nieuregulowanego za uregulowany.

Błąd standardowy średniej (*std. mean error*) — mówi o precyzji oszacowania średniej w populacji na podstawie średniej z próby. Im mniejsza wartość błędu, tym większa precyzja oszacowania. Zależy ona od dwóch wielkości:

— rozproszenia wartości zmiennej mierzonego za pomocą odchylenia standardowego: im większe rozproszenie wartości zmiennej w próbie, tym większy jest błąd standardowy;

— liczebności próby: im większa jest próba, tym mniejszy jest błąd standardowy; zależność ta nie ma jednak liniowego charakteru: aby uzyskać kolejny spadek wielkości błędu o pewien przedział wartości, trzeba coraz większych przyrostów wielkości próby.

Wartość błędu standardowego średniej uzyskuje się ze wzoru

$$d = \frac{s}{\sqrt{N}},$$

gdzie s oznacza odchylenie standardowe w próbie, a N — liczebność próby.

Błąd standardowy służy do obliczenia przedziału ufności dla średniej.

Błąd systematyczny Δx_{syst} (*systematic error, bias*) — błąd, który jest w pewnym sensie obciążony, tzn. ma rozkład prawdopodobieństwa z wartością oczekiwaną lub inną miarą położenia, różną od zera:

$$\Delta x_{syst} = \mu'_x - \mu_x,$$

gdzie: μ'_x — wartość średnia w zbiorze wyników; μ_x — wartość „prawdziwa" badanej cechy.

Błąd względny $\delta(a)$ (*relative error*) — liczby przybliżonej a, określa się jako stosunek błędu bezwzględnego $\Delta(a)$ tej liczby do wartości bezwzględnej odpowiedniej liczby dokładnej A ($A \neq 0$):

$$\delta(a) = \frac{\Delta(a)}{|A|}.$$

Certyfikowany materiał odniesienia (*certified reference material, CRM*) — materiał odniesienia opatrzony certyfikatem, charakteryzujący się wartością lub wartościami danej właściwości, które certyfikowano zgodnie z procedurą zapewniającą odniesienie do dokładnej realizacji jednostki miary, w której wyrażane są wartości danej właściwości; każdej wartości certyfikowanej powinna być przy tym przypisana niepewność odpowiadająca określonemu poziomowi ufności.

Częstość (*frequency*) — liczba realizacji zdarzenia określonego typu lub liczba elementów pewnej populacji, które należą do określonej klasy.

Częstość względna (*relative frequency*) — częstość indywidualnej grupy obserwacji w rozkładzie liczebności wyrażona w postaci frakcji ogólnej częstości.

Czułość metody analitycznej (*sensitivity*) — stosunek przyrostu sygnału analitycznego do odpowiadającego mu przyrostu stężenia lub zawartości oznaczanego składnika lub najmniejsza różnica między dwoma sygnałami na wejściu metody, która wywołuje dostrzegalną różnicę między odpowiednimi sygnałami na wyjściu metody.

Czułość metody analitycznej określona jest nachyleniem krzywej kalibracji tej metody. Jeśli krzywa ta nie jest linią prostą, czułość zmienia się wraz ze zmianą stężenia oznaczanej substancji. W przypadku gdy czułość jest funkcją składu matrycy, prosta kalibracja za pomocą czystych substancji nie jest adekwatna. Jeśli odpowiedź urządzenia jest liniowo zależna od stężenia oznaczanej substancji, należy stosować metodę dodatku wzorca.

Detekcja (*detection*) — wykrywanie obecności substancji z zastosowaniem reakcji chemicznych lub procesów fizycznych.

Dokładność (*accuracy*) — w sensie statystycznym różnica pomiędzy obliczonymi lub oszacowanymi i dokładnymi lub prawdziwymi wartościami; dokładność ocenia zgodność wyników uzyskiwanych dzięki stosowanej metodzie analitycznej z wartością przyjętą jako prawdziwą. W próbach modelowych tworzonych w laboratorium wartość prawdziwa to wartość odważona, w preparatach wykonanych zgodnie z zasadami Dobrej Praktyki Produkcyjnej (GMP, *Good Manufacturing Practice*), to deklaracja.

Dystrybuanta $F(x)$ (*cumulative distribution function*) — jej wartościami (otrzymuje się je przez kumulowanie wartości funkcji prawdopodobieństwa) są prawdopodobieństwa zdarzeń losowych

polegających na tym, że zmienna losowa (X) przybierze wartość mniejszą niż dana liczba rzeczywista x:

$$F(x) = P(X < x).$$

Funkcja aktywacji (*activation function*) — element neuronu w sieci MLP, funkcja nieliniowa wykorzystywana do modyfikowania sumowanych wejść neuronów.

Efekt matrycy (*matrix effect*) — zakłócenia w wykrywaniu lub/i oznaczaniu analitu spowodowane wpływem innych składników próbki.

Główny zbiornik wód podziemnych, GZWP (*major groundwater basin*) — zbiornik wód podziemnych odpowiadający umownie ustalonym ilościowym i jakościowym kryteriom podstawowym: wydajność potencjalnego otworu studziennego powyżej $70\,m^3 \cdot h^{-1}$, wydajność ujęcia powyżej $10000\,m^3 \cdot d^{-1}$, przewodność warstwy wodonośnej wyższa niż $10\,m^2 \cdot h^{-1}$, najwyższa klasa jakości wody. W obszarach deficytowych do wyznaczenia GZWP stosuje się indywidualne kryteria ilościowe. W Polsce wydzielono 180 GZWP (40 wg indywidualnych kryteriów) o łącznej powierzchni $163441\,km^2$ i szacunkowych zasobach dyspozycyjnych $7.35\,km^3 \cdot a^{-1}$.

Granica decyzji L_C (*limit of decision*) — granica, powyżej której można stwierdzić, czy wynik analizy wskazuje na obecność danego składnika w próbce.

Granica oznaczalności L_Q (*limit of determination*) — najmniejsze stężenie analitu w próbce, które może być dokładnie oznaczone.

Granicę tę (wyrażoną jako stężenie lub ilość), najmniejszy sygnał danej metody analitycznej, jaki może być oznaczony z zadowalającą pewnością, uzyskuje się ze wzoru:

$$L_Q = DL = \bar{x}_l + k\sigma_l,$$

gdzie: $L_Q = DL$ — granica oznaczalności; $\bar{x}_l$ — wartość średnia z wyników oznaczeń próbek ślepych; σ_l — oszacowane odchylenie standardowe wyników oznaczeń próbek ślepych; k — numeryczny współczynnik odpowiadający żądanemu poziomowi ufności. Zaleca się wykonywanie obliczeń dla $k = 6$.

Granica wykrywalności L_D (*limit of detection*) — określona jest jako najmniejsze stężenie analitu w próbce, które może być wykryte, niekoniecznie oznaczone (w danych warunkach eksperymentalnych).

Granicę tę (wyrażoną jako stężenie lub ilość), najmniejszy sygnał danej metody analitycznej, jaki może być wykryty z zadowalającą pewnością, uzyskuje się ze wzoru:

$$L_D = \bar{x}_l + k\sigma_l,$$

gdzie: L_D — granica wykrywalności; $\bar{x}_l$ — wartość średnia z wyników oznaczeń próbek ślepych; σ_l — oszacowane odchylenie standardowe wyników oznaczeń próbek ślepych; k — numeryczny współczynnik odpowiadający żądanemu poziomowi ufności. Zaleca się wykonywanie obliczeń dla $k = 3$.

Jakość wody (*water quality*) — właściwość wody opisana zespołem cech (wskaźników fizykochemicznych) stanowiących o przydatności wody do określonych celów.

Karta kontrolna (*control chart*) — dokument stosowany do statystycznej kontroli stabilności procesu produkcyjnego lub parametrów metody analitycznej, na którym są rejestrowane wyniki badania próbek pobieranych systematycznie z bieżącej produkcji, czy też (w przypadku metod analitycznych) wyników pomiarów badanych parametrów, stężeń analizowanego składnika. Szczegóły dotyczące sposobu tworzenia i analizy kart kontrolnych można znaleźć w literaturze — Szczepańska, Kmiecik, 1998.

Klasyfikacja (*classification problem*) — przypisanie przykładu do jednej z kilku dyskretnych kategorii, np. klasyfikacja punktu monitoringowego do określonej klasy zagrożenia wód, czy obszaru o określonym zagospodarowaniu terenu.

Kontrola jakości (*quality inspection/control*) — czynność lub zespół czynności mających na celu sprawdzenie zgodności właściwości produktu z wymaganiami.

Korelacja (*correlation*) — wzajemna zależność dwóch lub więcej zmiennych losowych taka, że jedna ze zmiennych reaguje zmianami swej wartości oczekiwanej na zmiany innych.

Kowariancja (*covariance*) — opis numeryczny tendencji kojarzenia się w pary wartości dużej z dużą, małej z małą.

Kurtoza (*kurtosis*) — zwana także współczynnikiem koncentracji (skupienia), służy do porównywania rozkładów liczności dwóch albo więcej cech o różnych mianach ze względu na skupienie, umożliwia badanie stopnia „spiczastości" pomiarów. Może być liczona ze wzoru

$$\text{kurtoza} = \frac{n(n+1)\sum_{i=1}^{n}(x_i-\bar{x})^4}{(n-1)(n-2)(n-3)s_x^4} - \frac{3(n-1)^2}{(n-2)(n-3)},$$

gdzie: n — liczba pomiarów; x_i — i-ty pomiar; $\bar{x}$ — wartość średnia uzyskanych pomiarów; s_x — odchylenie standardowe pomiarów.

Rozkład normalny, symetryczny (*mesokurtic*) ma współczynnik kurtozy równy 0. Jeśli współczynnik kurtozy ma wartość dodatnią, oznacza to spiczastość rozkładu (*leptokurtic*, gdy ma on wartość ujemną, rozkład jest spłaszczony (*platykurtic*).

Kurtoza standaryzowana (*std. kurtosis*) —

$$\text{stand. kurtoza} = \frac{\text{kurtoza}}{\sqrt{\frac{24}{n}}},$$

gdzie n to liczba pomiarów.

Kwantyl (*quantile*) — kwantylem rzędu p ($0 < p < 1$) zmiennej losowej X (lub jej rozkładu) nazywamy liczbę z_p, taką, że prawdopodobieństwo P:

$$P(X \leqslant z_p) \geqslant p, \quad P(X \geqslant z_p) \geqslant 1-p,$$

czyli, że

$$F(z_p) \leq p \leq F(z_p + 0),$$

gdzie $F(x)$ jest dystrybuantą zmiennej losowej X.

Każda zmienna losowa X ma kwantyl dowolnego rzędu p, ale nie zawsze kwantyl z_p jest określony jednoznacznie. **Mediana** jest kwantylem rzędu $\frac{1}{2}$. Kwantyl $z_{1/4}$ rzędu $\frac{1}{4}$ nosi nazwę **dolnego kwartyla**, a kwantyl $z_{3/4}$ rzędu $\frac{3}{4}$ — **górnego kwartyla**. Można je traktować jako miarę rozproszenia zmiennej losowej X.

Kwartyl (*quartile*) — patrz: **Kwantyl**.

Laboratorium akredytowane (*accredited laboratory*) — to laboratorium badawcze, któremu została udzielona akredytacja.

Laboratorium badawcze, pomiarowe (*testing laboratory*) — laboratorium wykonujące badania.

Laboratorium kalibrujące, wzorcujące (*calibration laboratory*) — laboratorium wykonujące wzorcowania.

Liczba stopni swobody (*number of degrees of freedom, number of independent variables*) — oznacza liczbę niezależnych obserwacji, których można użyć do obliczania danej charakterystyki; jeśli mamy n wyników pomiarów, to liczba stopni swobody wynosi $n-1$.

Linia centralna (*central line*) — linia na karcie kontrolnej, odpowiadająca wartości oczekiwanej kontrolowanego parametru statystycznego, obliczonej na podstawie badań wstępnych lub przyjętej zgodnie z określonymi założeniami teoretycznymi. Wartość ta jest nazywana również wartością normalną.

Linie kontrolne (*control lines*) — linie określające dopuszczalne odchylenia badanego parametru statystycznego od wartości oczekiwanej, służące do wnioskowania o jakości produktu na podstawie rozmieszczenia punktów w stosunku do położenia tych linii.

Liniowość (*linearity*) — w danym zakresie, jest to możliwość uzyskiwania wyników oznaczeń wprost proporcjonalnych do zawartości substancji oznaczanej w próbce. Miarą liniowości jest wariancja nachylenia krzywej regresji.

Macierz odwołań (*matrix of references*) — tabela, w której kolumny reprezentują klasę prognozowaną, a wiersze — klasę prawdziwą. Jeśli zgodność wyników prognozowanych z prawdziwymi jest 100%, wpisy znajdą się jedynie na przekątnej macierzy, a suma wpisów będzie równa liczbie obserwacji w zbiorze danych. W zagadnieniach decyzyjnych macierz korelacji jest zdefiniowana za pomocą symboli przypisanych do decyzji. W zagadnieniach predykcji macierz jest zdefiniowana za pomocą małych przedziałów wartości (np. wszystkie prognozy lub wyniki z przedziałów 0.1—0.3, 0.3—0.5, 0.5—0.7 itd. będą widziane „razem"). Można zmieniać liczbę tych przedziałów (*bins*).

Materiał odniesienia (*reference material, RM*) — materiał lub substancja, których jedna lub więcej wartości ich właściwości są dostatecznie jednorodne i na tyle dobrze określone, aby mogły być stosowane do wzorcowania przyrządu, do oceny metody pomiarowej lub do przypisania wartości właściwościom materiałów. Materiał odniesienia może być ciałem czystym lub mieszaniną i występować pod postacią gazu, cieczy lub ciała stałego. Przykładami są: woda do wzorcowania lepkościomierzy, szafir pozwalający wzorcować pojemność cieplną w kalorymetrii i roztwory wzorcowe stosowane do wzorcowania w analizie chemicznej.

Materiał odniesienia certyfikowany (*certified reference material, CRM*) — materiał odniesienia opatrzony certyfikatem, charakteryzujący się wartością lub wartościami danej właściwości, które certyfikowano zgodnie z procedurą zapewniającą odniesienie do dokładnej realizacji jednostki miary, w której wyrażane są wartości danej właściwości; każdej wartości certyfikowanej powinna być przy tym przypisana niepewność odpowiadająca określonemu poziomowi ufności.

Metoda analityczna (*analytical method*) — wszystkie czynności związane z wykonaniem analizy, począwszy od przygotowania badanej próbki, a skończywszy na pomiarze wielkości mierzonej będącej podstawą oznaczenia oraz opracowanie wyników. „Katalog wybranych fizycznych i chemicznych wskaźników zanieczyszczeń wód podziemnych" (Witczak, Adamczyk, 1994) oraz „Prawo ochrony środowiska..." (1996) zawierają wykaz metod analitycznych zalecanych dla potrzeb monitoringu wód podziemnych.

Metoda badania (*method of testing, research method*) — ustalona procedura techniczna wykonywania badań.

Metodyka opróbowania wód podziemnych (*methodology of groundwater sampling*) — najważniejszy element monitoringu jakości wód podziemnych, decydujący o jakości uzyskanych rezultatów. PIOŚ (Witczak, Adamczyk, 1994) określa szczegółowo sposób i zasady pobierania próbek dla potrzeb monitoringu wód podziemnych, WHO (1998) podaje zasady opróbowania wód przeznaczonych do picia.

Miara położenia (*measure of position*) — charakterystyka liczbowa zmiennej losowej, która po dodaniu do zmiennej losowej dowolnej stałej zmienia swą wartość o tę stałą (wartość oczekiwana, kwantyle, moda, mediana).

Miara rozrzutu (*measure of dispersion*) — charakterystyka liczbowa zmiennej losowej, która po dodaniu do zmiennej losowej dowolnej stałej nie zmienia swej wartości (wariancja, odchylenie standardowe, rozstęp).

Monitoring jakości wód podziemnych (*groundwater quality monitoring*) — kontrolno-decyzyjny system oceny dynamiki antropogenicznych przemian w wodach podziemnych. Polega na prowadzeniu w wybranych charakterystycznych punktach powtarzalnych pomiarów i badań stanu zwierciadła wód podziemnych i ich jakości oraz interpretacji ich wyników w aspekcie ochrony środowiska wodnego. Celem monitorowania wód podziemnych jest wspomaganie działań zmierzających do likwidacji lub ograniczenia ujemnego wpływu czynników antropogenicznych na wody podziemne.

W Polsce monitoring wód podziemnych jest prowadzony w sieciach: krajowej, regionalnych i lokalnych. Sieć krajową tworzą wybrane, reprezentatywne punkty obserwacyjne (aktualnie 726). Głównym zadaniem monitoringu regionalnego jest rozpoznanie oraz stała kontrola jakości wód w zbiornikach o znaczeniu regionalnym, w tym GZWP. Zadaniem monitoringu lokalnego jest rozpoznanie i śledzenie wpływu (stwierdzonych i potencjalnych) ognisk zanieczyszczeń na jakość wód podziemnych.

Niepewność pomiaru (*uncertainty of measurement*) — parametr związany z wynikiem pomiaru, charakteryzujący rozrzut wartości, które można w uzasadniony sposób przypisać wielkości mierzonej.

Norma (*standard*) — przyjęta, obowiązująca, wymagana miara, ilość, jakość czegoś.

Obserwacje (*observations*) — wyniki badania próbki na określoną jej cechę.

Obszar typowy zmiennej (*typical variable area*) — obszar pomiędzy pierwszym i trzecim kwartylem, w którym leży 50% obserwacji.

Odchylenie standardowe w próbce (*std. deviation*) — dodatni pierwiastek kwadratowy z wariancji w próbce:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

gdzie: n — liczba pomiarów; x_i — i-ty pomiar; $\bar{x}$ — wartość średnia uzyskanych pomiarów.

Odtwarzalność wyników pomiarów (*reproducibility*) — stopień zgodności wyników pomiarów tej samej wielkości mierzonej, wykonywanych w zmienionych warunkach pomiarowych; warunki podlegające zmianom mogą obejmować: zasadę pomiaru, metodę pomiaru, obserwatora, przyrząd pomiarowy, etalon odniesienia, miejsce, warunki stosowania, czas.

Opróbowanie wód podziemnych (*groundwater sampling*) — czynności związane z metodycznym pobieraniem próbek wód podziemnych. Do opróbowania zalicza się także pomiary głębokości zwierciadła wody, temperatury i niektóre polowe oznaczenia jej właściwości.

Parametr próbki, statystyka (*statistics*) — jednoznacznie określona funkcja obserwacji z próbki, np. wartość średnia, suma wszystkich wartości wyników badania próbki itp.

Parametr statystyczny (*statistical parameter*) — patrz: **Parametr próbki**.

Parametry kontroli jakości (*data quality parameters*) — według Nielsena (1991) to precyzja, dokładność, reprezentatywność, porównywalność i kompletność. USP XXIII i IUPAC wymieniają jeszcze granice: decyzji, wykrywalności i oznaczalności, liniowość, rozstęp oraz selektywność i specyficzność.

Percentyle (*percentiles*) — zbiór wartości dzielących ogólną liczebność na sto równych części.

Pobieranie próbek (*sampling*) — czynność lub zespół czynności, związanych z wyznaczaniem i pobieraniem z partii (lub procesu technologicznego) tych jednostek produktu, które będą stanowić próbkę.

Pobieranie próbek losowe (*random sampling*) — pobieranie próbek w sposób zapewniający każdej jednostce produktu w partii (lub w procesie technologicznym) jednakowe prawdopodobieństwo znalezienia się w próbce.

Pomiar (*measurement*) — zbiór operacji mających na celu wyznaczenie wartości wielkości.

Populacja (*population*) — zbiór jednostek jednego rodzaju (osobników) lub zbiór obserwacji (wyniki ważenia, miareczkowania, innego rodzaju pomiary); dla podkreślenia, że chodzi o całość badanego zjawiska, używa się określenia populacja generalna (patrz: **Populacja generalna**).

Populacja generalna, zbiorowość generalna (*parent population*) — zbiór posiadający pewną właściwość wspólną dla wszystkich jego części, kwalifikującą je do tego zbioru oraz przynajmniej jedną właściwość (cechę), ze względu na którą części populacji (populacje, podzbiory) mogą się między sobą różnić.

Powtarzalność wyników pomiarów (*repeatability*) — stopień zgodności wyników pomiarów tej samej wielkości mierzonej, wykonywanych w tych samych warunkach pomiarowych (ta sama procedura pomiarowa, ten sam operator, ten sam przyrząd pomiarowy stosowany w tych samych warunkach, to samo miejsce, krótkie odstępy czasu).

Poziom ufności (*level of confidence*) — prawdopodobieństwo zdarzenia, że przedział ufności pokryje nieznaną wartość parametru populacji.

Praktyczna granica oznaczalności (*practical quantification limit*) — ma taki sam sens jak granica oznaczalności, liczona z takiego samego wzoru jednak zamiast próbek ślepych bada tzw. próbki zerowe.

Precyzja (*precision*) — miara zgodności między wartościami eksperymentalnymi otrzymanymi w ciągu całości badań wykonanych w określonych warunkach.

Precyzja metody analitycznej (*precision of analytical method*) — wielkość charakteryzująca rozrzut wyników uzyskiwanych przy wielokrotnym oznaczaniu danego składnika daną metodą, określona odchyleniem standardowym i rozstępem wyników uzyskiwanych w warunkach powtarzalności lub odtwarzalności.

Predykcja (*prediction*) — polega na przypisaniu badanemu przypadkowi określonej wartości liczbowej, np. prognozowanie stężeń wskaźników fizyko-chemicznych wód na podstawie współrzędnych punktu monitoringowego.

Procesy hydrogeochemiczne (*hydrogeochemical processes*) — procesy współdziałania wód podziemnych z ośrodkiem skalnym, zmieniające w wymierny sposób chemizm i właściwości wód oraz równocześnie skład chemiczny skał.

Próbka (*sample*) — dowolny skończony podzbiór populacji, podlegający bezpośredniemu badaniu.

Próbki dublowane (*duplicate samples*) — pobierane losowo z wybranych punktów monitoringu jako duplikaty próbek normalnych. Próbki te służą do oceny precyzji uzyskiwanych wyników.

Próbki kontrolne (*control samples*) — próbki pobierane do specjalnych celów np. określania precyzji badań, wyznaczania granic oznaczalności, w monitoringu wód podziemnych są to próbki zerowe, dublowane i znaczone.

Próbki normalne (*normal samples*) — próbki pobierane w monitoringu jakości wód podziemnych.

Próbki ślepe (*blank samples*) — próbki wykorzystywane do obliczenia laboratoryjnych granic wykrywalności i oznaczalności. Ślepa próbka — roztwór stanowiący wodę zdejonizowaną z matrycą w postaci tych samych odczynników, które zawierają roztwory wzorcowe. Ślepa próbka jest w czasie analizy poddawana identycznej obróbce jak badana próbka, np. zagęszczanie, separacja.

Próbki zerowe (*zero samples*) — pobierane tym samym sprzętem co próbki normalne, ale z użyciem jako medium wody dejonizowanej o wysokiej czystości; odbywają taką samą obróbkę, transport i przechowywanie jak próbki normalne, służą do wyznaczenia praktycznej granicy oznaczalności PDL.

Próbki znaczone (*spiked samples*) — o znanym składzie lub dodatku wzorca wybranych substancji, pozwalają ocenić dokładność, a więc wykryć błędy losowe i ewentualny błąd systematyczny.

Przedział tolerancji (*tolerance interval*) — zbudowany na podstawie wyników badania próbki przedział, który z określonym prawdopodobieństwem P zawiera co najmniej $100Q\%$ populacji generalnej (Q — parametr populacji).

Przedział ufności (*confidence interval*) — zbudowany na podstawie badania próbki przedział, który z określonym prawdopodobieństwem pokrywa prawdziwą, nieznaną wartość parametru populacji.

Przedział ufności wyznacza się przy założonym współczynniku ufności, który mówi, jakie jest prawdopodobieństwo, że obliczony przedział wartości pokryje prawdziwą wartość średniej w populacji generalnej. Przy danym poziomie błędu standardowego, im większą wartość współczynnika ufności przyjmiemy, tym szerszy będzie przedział ufności. Oszacowanie będzie więc bardziej wiarygodne, ale mniej precyzyjne. Standardowo przyjmuje się wartość współczynnika ufności na poziomie 95%. Wówczas granice przedziału ufności wyznaczane są ze wzoru

$$\text{średnia z próby} \pm 1.96 \times \text{błąd standardowy}.$$

W przypadku przyjęcia 99% poziomu ufności, zamiast parametru 1.96 wstawiamy do wzoru współczynnik 2.58. Wielkości te wynikają z właściwości rozkładu normalnego, któremu podlegają średnie obliczone z wielu prób wylosowanych z danej populacji (centralne twierdzenie graniczne).

Przeuczenie sieci (*overtraining*) — jeśli proces uczenia trwa zbyt długo, sieć traci zdolność uogólniania.

Realizacja zmiennej losowej (*realization of random variable*) — wartość zmiennej losowej dla konkretnego zdarzenia elementarnego.

Rozkład prawdopodobieństwa zmiennej losowej (*probability distribution of random variable*) — reguła, według której wartości prawdopodobieństwa są przypisane wartościom zmiennej losowej;

najczęściej wykorzystywane typy rozkładów są zebrane i przedstawione w pracach m.in. (Benjamin, Cornell, 1977; Czermiński et al., 1994; Kmiecik, 1995).

Rozstęp w próbce (*range*) — różnica między wartościami największą i najmniejszą w próbce; rozstęp w próbce o liczności n jest różnicą między n-tą ($x_{(n)}$) i pierwszą ($x_{(1)}$) statystyką pozycyjną:

$$R = x_{(n)} - x_{(1)}$$

patrz: **Rozstęp ruchomy**.

Rozstęp ruchomy (*moving range*) — przy założeniu istnienia przedziałów ruchomych rzędu q, gdzie $q = 1, 2, \ldots, N$ (należy pamiętać, że $\max N = 25$) sztucznie tworzy się dwu- lub trzyelementowe próbki (tzw. „pseudopróbki") i oblicza tzw. rozstępy ruchome. Mając N pojedynczych wyników x_1, x_2, ..., x_N, i-ty rozstęp ruchomy określa się jako różnicę między największą i najmniejszą wartością w i-tej pseudopróbce. Z kolei, i-ta pseudopróbka o liczności n składa się z elementów x_i, x_{i+1}, ..., x_{i+n-1} (gdzie $i = 1, 2, \ldots, N - n + 1$). Liczebność pseudopróbki n ustala się zwykle równą 2 lub 3. Zatem dla $n = 2$, i-ty rozstęp ruchomy oznaczamy symbolem MR_i.

$$MR_1 = |x_2 - x_1|,\ MR_2 = |x_3 - x_2|,\ MR_{N-1} = |x_N - x_{N-1}|.$$

Rozwiązanie globalne (*global solution*) — najlepszy zbudowany model sieci neuronowej, definiowany jako rozwiązanie z najmniejszym błędem.

Różnica bezwzględna (*absolute difference*) — wartość bezwzględna różnicy między wartościami dwu zmiennych, w szczególności dwu zmiennych losowych.

Segmentacja (*segmentation*) — podział danych na grupy o podobnych cechach.

Seria punktów na karcie kontrolnej (*run*) — zbiór kolejnych punktów o ustalonych właściwościach, który jest poprzedzony, i który poprzedza zbiory punktów o innych właściwościach. Serią punktów na karcie kontrolnej jest np. zbiór kolejnych punktów powyżej linii centralnej, poprzedzony zbiorem punktów poniżej tej linii, po którym następuje zbiór punktów poniżej linii centralnej. Liczbę kolejnych punktów w serii nazywa się długością serii.

Siatka probabilistyczna (*probability paper*) — papier z odpowiednio przygotowaną skalą, na którym dystrybuanta danego rozkładu może być wykreślona w postaci linii prostej, jeżeli na osi odciętych odkładane są wartości zmiennej losowej.

Sieć monitoringu jakości wód podziemnych (*groundwater-monitoring network*) — sieć otworów hydrogeologicznych, w których są dokonywane systematycznie pomiary zmian położenia zwierciadła wód podziemnych i badania jakości. Może mieć charakter stały lub okresowy (zadaniowy, celowy).

Standaryzacja zmiennej (*standardization* — bardzo często porównanie rozkładów dwóch lub więcej zmiennych czy też konstruowanie indeksów opartych na sumowaniu wartości kilku zmiennych wymaga wyeliminowania wpływu jednostek miary na rozkład zmiennej, przy zachowaniu wzajemnej proporcji wartości, służy temu standaryzacja zmiennych.

Klasyczny rodzaj standaryzacji to tzw. **standaryzacja empiryczna**, albo standaryzacja typu Z, która polega na odjęciu od każdej wartości zmiennej jej średniej i podzieleniu wyniku przez odchylenie standardowe:

$$Z_i = \frac{X_i - \bar{X}}{s}.$$

Standaryzowany współczynnik asymetrii (*std. skewness*) —

$$\text{stand. współcz. asym.} = \frac{\text{współcz. asym.}}{\sqrt{\frac{6}{n}}},$$

gdzie n to liczba pomiarów.

Statystyczna Kontrola Jakości (*Statistical Quality Inspection/Control*) — kontrola wyrywkowa, w której są stosowane metody statystyczne do wnioskowania o jakości partii produktu lub stabilności procesu technologicznego na podstawie wyników badania jednej lub wielu próbek.

Statystyka pozycyjna (*positional statistics*) — rzędu k jest to zmienna losowa, która przyjmuje k-tą co do wielkości wartość w uporządkowanej — według wartości niemalejących — próbce o liczności n; dla każdej próbki o liczności n możemy utworzyć n statystyk pozycyjnych.

Statystyki opisowe (*descriptive statistics*) — metody służące do organizacji, opisu i syntetycznej prezentacji danych liczbowych dotyczących pewnej zbiorowości.

Sygnał (*signal*) — takie rozmieszczenie punktów na torze karty kontrolnej w stosunku do linii kontrolnych, które uznaje się za dostatecznie pewny dowód rozregulowania procesu.

Sygnał pojedynczy (*single signal*) — sygnał polegający na wystąpieniu jednego punktu poza jedną z linii kontrolnych.

Sygnał seryjny, sekwencyjny (*sequential signal*) — sygnał polegający na wystąpieniu dwóch lub więcej kolejnych punktów poza tę samą linię kontrolną.

Sygnał uprzedzający (*attentive signal*) — sygnał polegający na wystąpieniu jednego punktu w obszarze między wewnętrzną i zewnętrzną linią kontrolną.

Średnia obcięta (*cut mean*) — wartość średnia obliczona po odrzuceniu obserwacji skrajnych, często odrzuca się 5% przypadków z góry i z dołu bazy danych, posortowanej wcześniej według wartości analizowanej zmiennej (5% średnia obcięta).

Im jej wartość jest bliższa wartości średniej, tym mniejszy wpływ na wartość średnią mają wartości odstające.

Tło hydrogeochemiczne (*hydrogeochemical background*) — zakres stężeń badanych substancji lub zakres wartości cech hydrochemicznych, charakterystyczny dla badanego środowiska, jednostki lub fragmentu jednostki hydrogeologicznej jednolitej pod względem hydrogeochemicznym.

Tło hydrogeochemiczne jest ograniczone dolną i górną granicą (wartości stężeń), poza którymi występują wartości anomalne.

Rozróżnia się **tło hydrogeochemiczne ogólne**, obejmujące zespół badanych substancji i cech hydrogeochemicznych oraz **tła cząstkowe** — dotyczące jednej cechy, np. tło chlorkowe.

Używa się również pojęcia **tła hydrogeochemicznego regionalnego** i **tła hydrogeochemicznego lokalnego**. Wyróżnia się też m.in. **tło hydrogeochemiczne pierwotne** (naturalne) oraz **tło hydrogeochemiczne współczesne**.

Tor karty kontrolnej (*track of control chart*) — część karty kontrolnej z układem osi o odciętej przedstawiającej kolejny numer próbki i rzędnej, przedstawiającej wartość badanego parametru statystycznego.

Uczenie nadzorowane (*supervised learning*) — proces uczenia sieci z wykorzystaniem mechanizmu uczącego, który przypisuje wartości docelowe do danych wejściowych.

Uczenie nienadzorowane (*unsupervised learning*) — proces uczenia sieci bez mechanizmu uczącego.

Układ wielkości (*system of quantities*) — zbiór wielkości, w znaczeniu ogólnym, między którymi istnieją określone relacje.

Wagi połączeń (*weights*) — wartości przypisane do połączeń między neuronami w sieci neuronowej; w trakcie uczenia się sieci wagi te są dopasowywane.

Wariancja w próbce (*variance*) — miara rozrzutu (rozproszenia) wartości zmiennej losowej od wartości oczekiwanej tej zmiennej (duża wartość wariancji świadczy o dużym rozproszeniu):

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

gdzie: n — liczba pomiarów, x_i — i-ty pomiar, $\bar{x}$ — wartość średnia uzyskanych pomiarów.

Wariancja analityczna (*analytical variance*) — zmienność związana z analityką wykonywanych pomiarów.

Wariancja całkowita (*total variance*) — suma σ_g^2 — wariancji hydrogeochemicznej, σ_s^2 — wariancji opróbowania, i σ_a^2 — wariancji analitycznej:

$$\sigma_{tot}^2 = \sigma_g^2 + \sigma_s^2 + \sigma_a^2.$$

Wariancja hydrogeochemiczna (*hydrogeochemical variance*) — związana ze średnim stężeniem analitu w terenie.

Wariancja opróbowania (*sampling variance*) — zmienność związana z błędami procesu opróbowania (systematycznym i przypadkowym).

Wariancja techniczna (*technical variance*) — suma efektów opróbowania σ_s^2 i analityki σ_a^2:

$$\sigma_{tech}^2 = \sigma_s^2 + \sigma_a^2.$$

Warstwa ukryta neuronów (*hidden layer*) — warstwa węzłów w sieci neuronowej, która nie ma bezpośrednich połączeń poza siecią.

Warstwy neuronów (*layers*) — grupy neuronów przetwarzanych równolegle; sieci neuronowe składają się z warstw neuronów.

Wartość modalna (*mode*) — wartość najczęstsza, moda, dominanta — wyznaczana jedynie z szeregu rozdzielczego.

Wartość oczekiwana, przeciętna (*average, mean*) — w przypadku zmiennej losowej dyskretnej definiowana jest następująco:

$$E(X) = \sum_i x_i p_i,$$

gdzie: x_i — i-ta obserwacja; p_i — prawdopodobieństwo wystąpienia danej obserwacji.

W przypadku zmiennej losowej ciągłej funkcję prawdopodobieństwa zastępujemy funkcją gęstości prawdopodobieństwa, zaś zabieg sumowania — całkowaniem:

$$E(X) = \int_a^b x \cdot f(x)dx.$$

Dla n równoważnych obserwacji (próba n-elementowa):

$$\bar{x} = \frac{1}{n}\sum x_i.$$

Wartość środkowa, mediana (*median*) — wartość zmierzona znajdująca się w środku uporządkowanego, rosnącego lub malejącego szeregu n wyników pomiarowych. W zależności od liczby obserwacji w tym szeregu medianę oblicza się następująco:

$$\tilde{x} = \begin{cases} x_{\frac{1}{2}(n+1)} & \text{dla nieparzystych wartości } n \\ \dfrac{x_{\frac{1}{2}n} + x_{\frac{1}{2}n+1}}{2} & \text{dla parzystych wartości } n. \end{cases}$$

Wielkość (*quantity*) — cecha zjawiska, ciała lub substancji, którą można wyróżnić jakościowo i wyznaczyć ilościowo.

Właściwość (*characteristic*) — wielkość, która daje się zmierzyć lub oszacować, i która może stanowić podstawę do różnicowania między sobą jednostek produktu w rozpatrywanej zbiorowości.

Współczynnik asymetrii (*skewness*) — zwany także współczynnikiem skośności, służy do porównywania rozkładów liczności dwóch albo więcej cech o różnych mianach ze względu na asymetrię. Oblicza się go na podstawie wzoru (podano wzory, jakie wykorzystywane są przez programy SPSS i QI Analyst do obliczania tych charakterystyk):

$$\text{współcz. asymetrii} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s_x^3},$$

gdzie: n — liczba pomiarów; x_i — i-ty pomiar; $\bar{x}$ — wartość średnia uzyskanych pomiarów; s_x — odchylenie standardowe pomiarów.

W przypadku rozkładu normalnego symetrycznego współczynnik asymetrii (skośności) ma wartość 0. Gdy występuje asymetria prawostronna (skośność dodatnia) współczynnik skośności ma wartość dodatnią, w przypadku asymetrii lewostronnej — wartość ujemną.

Wynik pomiaru (*result of a measurement*) — wartość przypisana wielkości mierzonej, uzyskana drogą pomiaru.

Wykres „łodyga i liście" (*steam-and-leaf*) — dzieli badany zbiór na klasy. „Łodygę" tworzą początkowe cyfry wartości zmiennej uporządkowane rosnąco i oddzielone od „liści" znakiem kropki. „Liście" tworzą kolejne cyfry zmiennej, przy czym każdy „liść" tworzą dwie pary cyfr, rozpoczynając od pary 0 i 1, a kończąc na parze 8 i 9. Jeżeli rozstęp między wartością minimalną i maksymalną jest duży, zakres „liścia" zwiększa się do pięciu kolejnych cyfr danego rzędu wielkości: od 0 do 4 i od 5 do 9. Długość „liścia" zależy od liczby wartości zmiennej, w których cyfry są identyczne.

Wykres skrzynkowy (*box-and-whisker plot*) — wykres opisowy tworzony na podstawie mediany, kwartyli i wartości skrajnych. Skrzynka reprezentuje rozstęp ćwiartkowy, który obejmuje 50% wartości. Wąsy są liniami rozciągającymi się od skrzynki do największej i najmniejszej wartości, wyłączając wartości odstające. Linia przechodząca przez skrzynkę wskazuje medianę.

Na podstawie wykresu typu „skrzynka z wąsami" można oceniać tendencję centralną (usytuowanie pionowej kreski wewnątrz pudełka na poziomie mediany), zróżnicowanie (wysokość pudełka, długość bocznych „wąsów" oraz kółka i krzyżyki poza „wąsami") oraz asymetrię rozkładu (odległość boków skrzynki od pionowek kreski i relacje długości „wąsów"). Kółeczka to obserwacje oddalone od boków pudełka o więcej niż półtora rozstępu kwartylnego, natomiast krzyżyki informują o oddaleniu większym niż 3 rozstępy kwartylne. Obserwacje te, jako zdecydowanie nietypowe powinny być usuwane ze zbioru danych (Luszniewicz, Słaby, 1997).

Wzorcowanie, kalibracja (*calibration*) — czynności, które wykonane w określonych warunkach pozwalają określić relację pomiędzy wartością wykazywaną przez urządzenie miernicze, system pomiarowy lub wynikającą z innego pomiaru, a znaną wielkością wzorca porównawczego.

Względne odchylenie standardowe, współczynnik zmienności (*factor of variability; relative standard deviation*) — określa się wzorem:

$$V = \frac{\sigma_x}{\bar{x}} \cdot 100\ [\%]$$

gdzie: $\bar{x}$ — wartość średnia pomiarów; σ_x — odchylenie standardowe.

Zapewnienie jakości/kontrola jakości QA/QC (*quality assurance/quality control*) — wszystkie planowane i systematyczne działania niezbędne do stworzenia odpowiedniego stopnia zaufania, że próbka, wyrób lub usługa spełni ustalone wymagania jakościowe.

Zbiorowość próbna (*sample*) — część populacji generalnej, na podstawie której wnioskuje się o właściwościach całej populacji.

Zmienna losowa (*random variable*) — funkcja, która przyporządkowuje wartości liczbowe zdarzeniom losowym będącym podzbiorem zbioru zdarzeń elementarnych.

Zmienna losowa ciągła (*continous random variable*) — przyjmuje wartości wyrażające się dowolnymi liczbami rzeczywistymi z określonych przedziałów (np. temperatura, stężenie roztworu, objętość, masa, współczynnik załamania światła, prędkość reakcji chemicznej, czas bezawaryjnej pracy danego urządzenia).

Zmienna losowa dyskretna (*discrete random variable*) — przybiera wartości wyrażające się tylko niektórymi liczbami rzeczywistymi, najczęściej całkowitymi, nieujemnymi (np. liczba cząstek jonizujących emitowanych przez określoną substancję w jednostce czasu, liczba wad technologicznych w jednej sztuce wyrobu, liczba sztuk wadliwych w partii towaru, liczba cykli pracy wyłącznika elektrycznego).

Przykład pełnej analizy statystycznej dla zbioru danych wejściowych

Analizowana zmienna:
cynk [mg/dm3] w wodach podziemnych sieci RMWP dorzecza górnej Wisły

Informacja o analizowanych danych

	Obserwacje					
	Uwzględnione		Wykluczone		Ogółem	
	N	Procent	N	Procent	N	Procent
Cynk [mg/dm3]	166	99.4%	1	.6%	167	100.0%

Statystyki opisowe

		Statystyka	Błąd standardowy
Średnia		.23149	9.9891E-02
95% przedział ufności dla średniej	Dolna granica	3.4264E-02	
	Górna granica	.42872	
5% średnia obcięta		7.0948E-02	
Mediana		4.0000E-02	
Wariancja		1.656	
Odchylenie standardowe		1.28701	
Minimum		.003	
Maksimum		15.000	
Rozstęp		14.997	
Rozstęp ćwiartkowy		7.5500E-02	
Skośność		10.174	.188
Kurtoza		110.748	.375

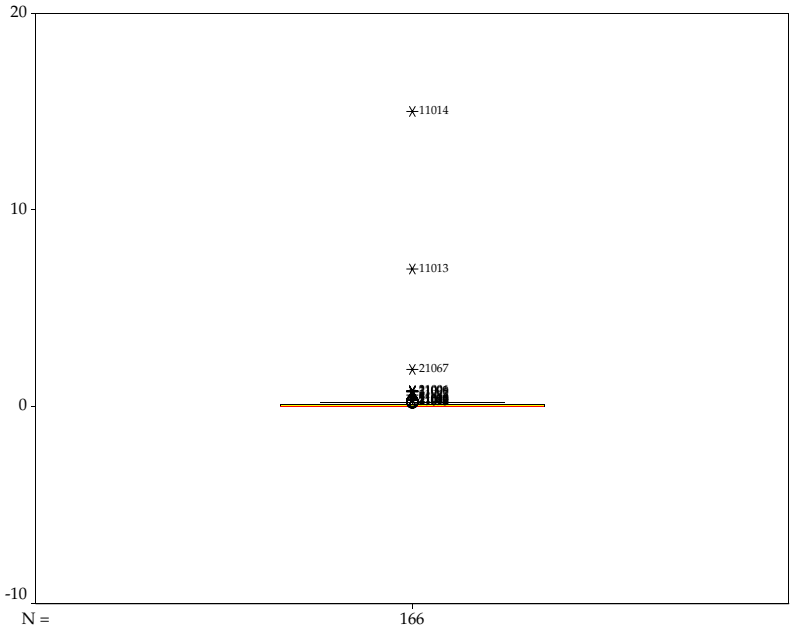
Percentyle

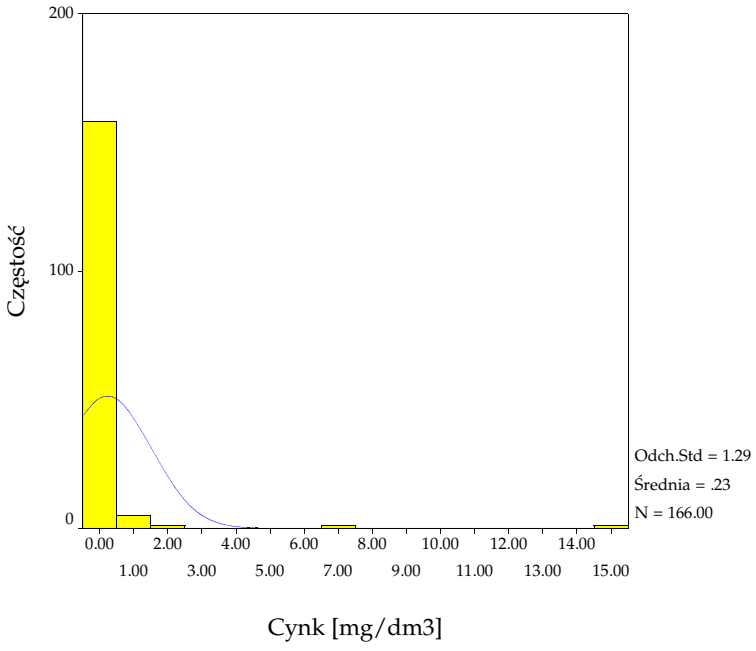
Percentyle	Przeciętne ważone (Definicja 1)	Zawiasy Tukey'a
5	9.3500E-03	
10	1.0000E-02	
25	2.0000E-02	2.0000E-02
50	4.0000E-02	4.0000E-02
75	9.5500E-02	9.5000E-02
90	.25470	
95	.52685	

Wartości skrajne

		Numer obserwacji	Numer id. punktu w bazie MONBADA	Wartość
Najwyższe	1	13	11014	15.000
	2	12	11013	7.000
	3	114	21067	1.900
	4	60	21006	.824
	5	1	11001	.800
Najniższe	1	152	21105	.003
	2	151	21104	.004
	3	72	21020	.004
	4	108	21061	.005
	5	96	21046	.006

```
Cynk [mg/dm3] Stem-and-Leaf Plot

 Frequency     Stem &  Leaf

       8.00      0 .  34456899
      23.00      1 .  00000000000000000012566699
      27.00      2 .  000000000000000000011123345678 8
      23.00      3 .  00000000000001112346677888
      12.00      4 .  000111233348
      10.00      5 .  00002235 88
       9.00      6 .  00000368 8
       8.00      7 .  00233569
       3.00      8 .  116
       3.00      9 .  057
       4.00     10 .  5667
       1.00     11 .  6
       4.00     12 .  0025
       1.00     13 .  5
       2.00     14 .  39
       1.00     15 .  8
       3.00     16 .  006
       2.00     17 .  07
        .00     18 .
       2.00     19 .  15
      20.00  Extremes    (>=.220)

 Stem width:        .010
 Each leaf:         1 case(s)
```

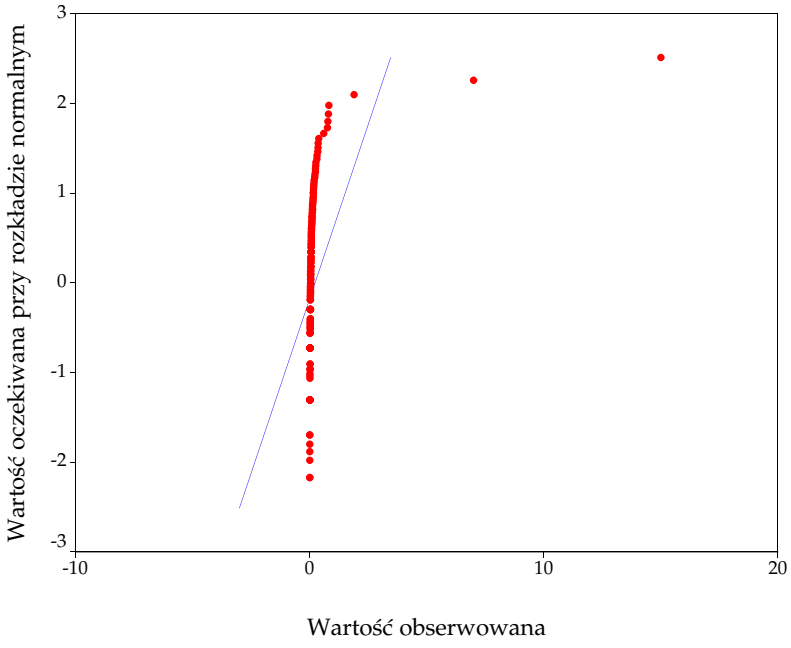

20
10
0
-10
11014
11013
21067
N =
166

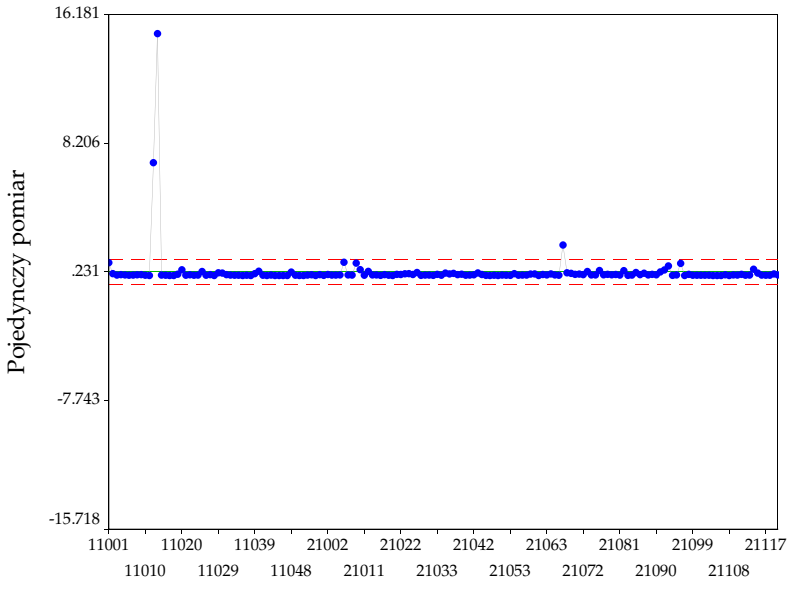

200
100
0
Częstość
0.00
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
9.00
10.00
11.00
12.00
13.00
14.00
15.00
Odch.Std = 1.29
Średnia = .23
N = 166.00
Cynk [mg/dm3]

Testy normalności rozkładu

	Kołmogorow-Smirnow[a]		
	Statystyka	df	Istotność
Cynk [mg/dm3]	.430	166	.000

a. Z poprawką istotności Lillieforsa





dodatek D

Przykład pełnej analizy statystycznej dla podzbioru wartości typowych

Analizowana zmienna:
cynk [mg/dm3] w wodach podziemnych sieci RMWP dorzecza górnej Wisły

Informacja o analizowanych danych

	Obserwacje					
	Uwzględnione		Wykluczone		Ogółem	
	N	Procent	N	Procent	N	Procent
Cynk [mg/dm3]	146	87.4%	21	12.6%	167	100.0%

Statystyki opisowe

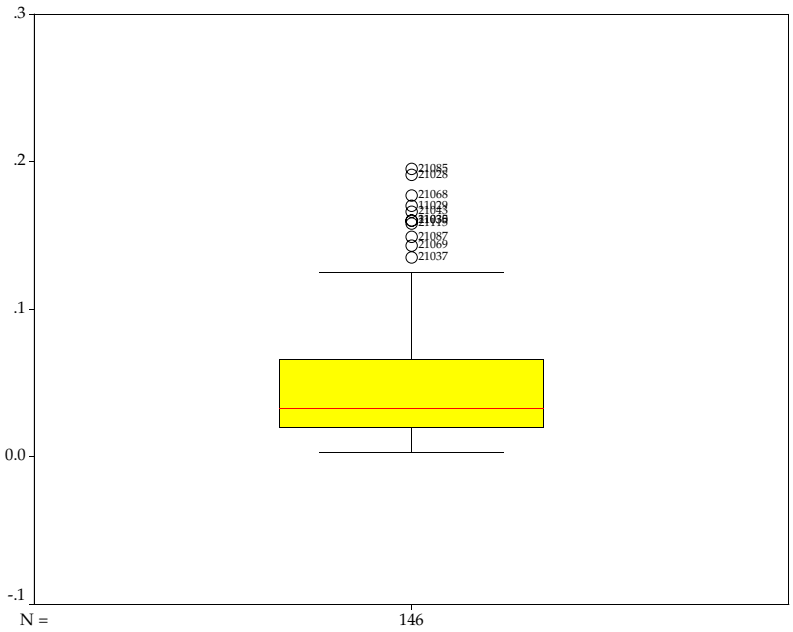
		Statystyka	Błąd standardowy
Średnia		4.9336E-02	3.6242E-03
95% przedział ufności dla średniej	Dolna granica	4.2173E-02	
	Górna granica	5.6499E-02	
5% średnia obcięta		4.4862E-02	
Mediana		3.2500E-02	
Wariancja		1.918E-03	
Odchylenie standardowe		4.3792E-02	
Minimum		.003	
Maksimum		.195	
Rozstęp		.192	
Rozstęp ćwiartkowy		4.6500E-02	
Skośność		1.522	.201
Kurtoza		1.764	.399

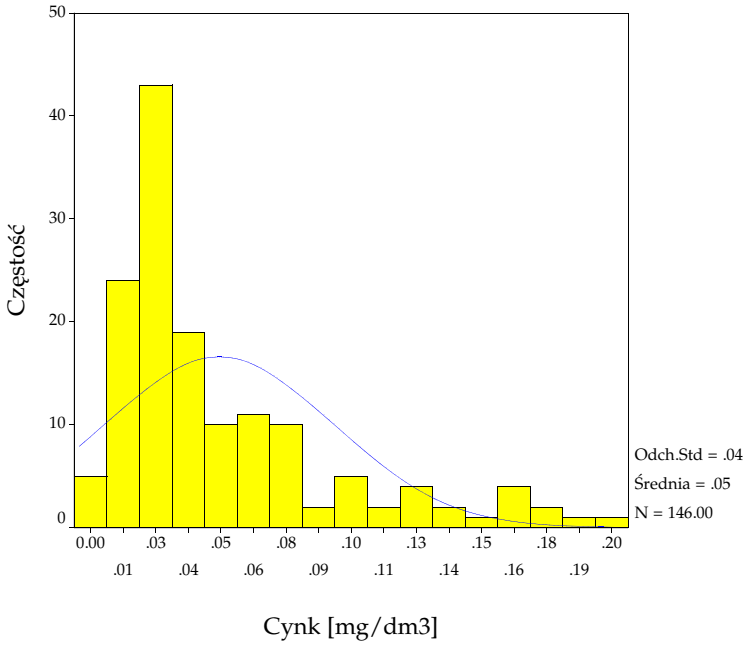
Percentyle

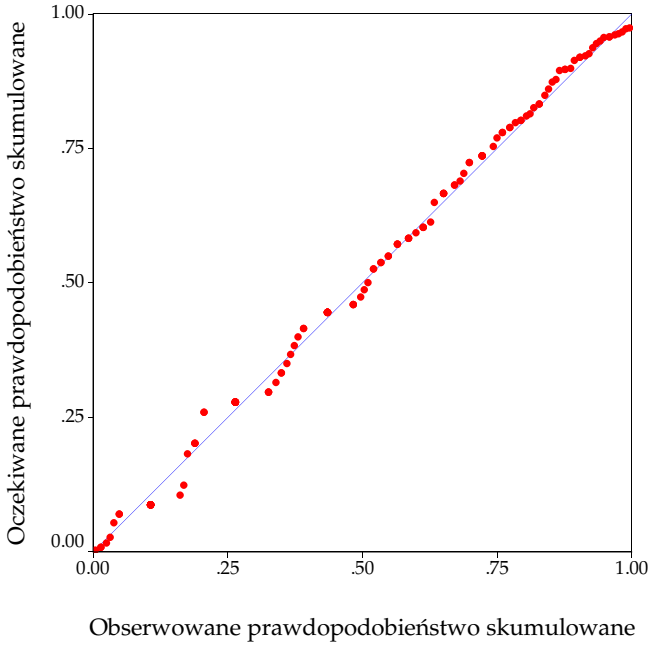
Percentyle	Przeciętne ważone (Definicja 1)	Zawiasy Tukey'a
5	9.0000E-03	
10	1.0000E-02	
15	1.0000E-02	
16	1.0520E-02	
25	2.0000E-02	2.0000E-02
50	3.2500E-02	3.2500E-02
75	6.6500E-02	6.6000E-02
84	8.7920E-02	
85	9.4750E-02	
90	.12000	
95	.15930	

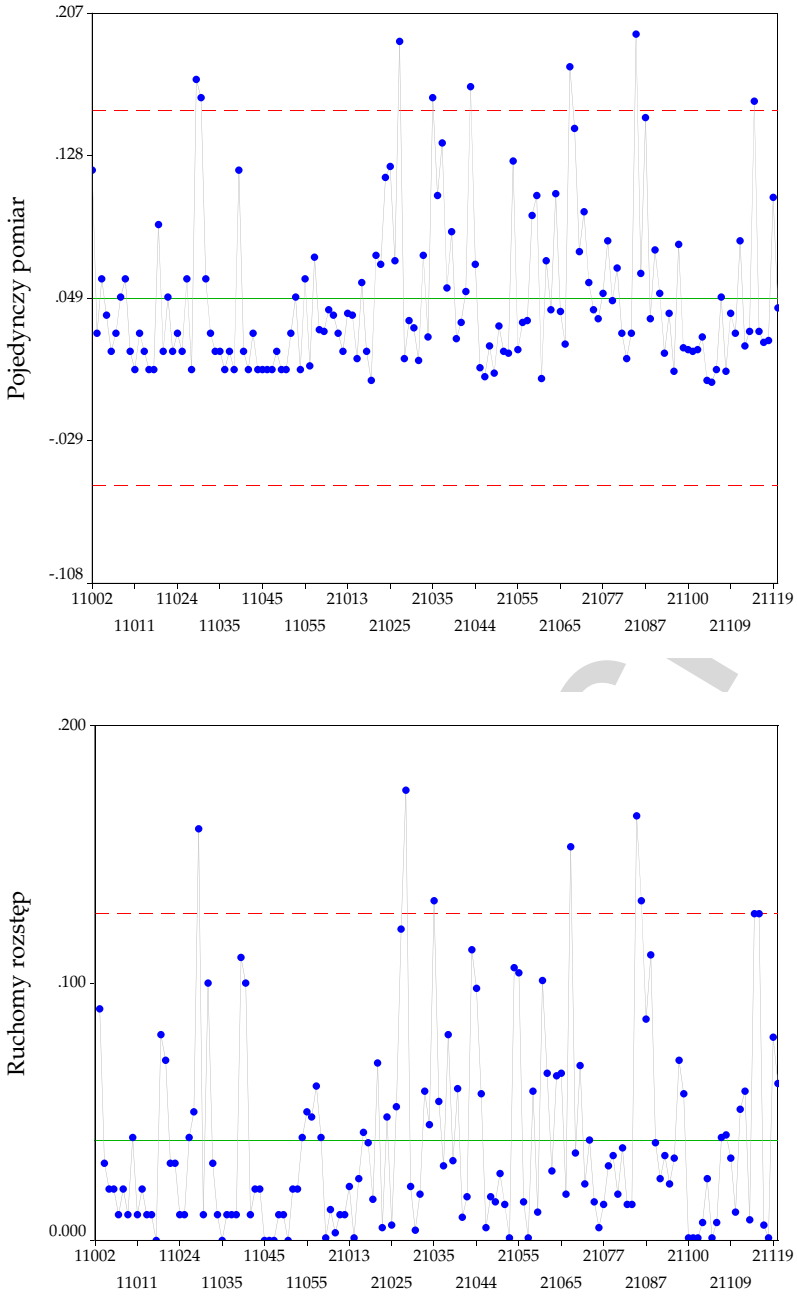
Wartości skrajne

		Numer obserwacji	Numer id. punktu w bazie MONBADA	Wartość
Najwyższe	1	132	21085	.195
	2	78	21028	.191
	3	115	21068	.177
	4	28	11029	.170
	5	93	21043	.166
Najniższe	1	152	21105	.003
	2	151	21104	.004
	3	72	21020	.004
	4	108	21061	.005
	5	96	21046	.006




Częstość
50
40
30
20
10
0
0.00
.01
.03
.04
.05
.06
.08
.09
.10
.11
.13
.14
.15
.16
.18
.19
.20
Odch.Std = .04
Średnia = .05
N = 146.00
Cynk [mg/dm3]


Oczekiwane prawdopodobieństwo skumulowane
1.00
.75
.50
.25
0.00
0.00
.25
.50
.75
1.00
Obserwowane prawdopodobieństwo skumulowane


Pojedynczy pomiar
.207
.128
.049
-.029
-.108
Ruchomy rozstęp
.200
.100
0.000
11002
11024
11045
21013
21035
21055
21077
21100
21119
11011
11035
11055
21025
21044
21065
21087
21109
Numer id. punktu w bazie MONBADA

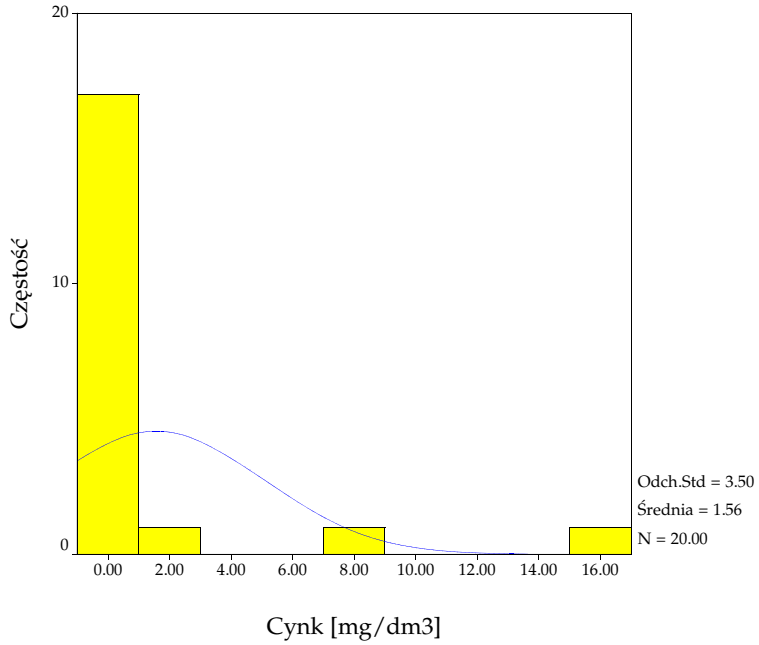
Przykład pełnej analizy statystycznej dla podzbioru wartości anomalnych

Analizowana zmienna:
cynk [mg/dm3] w wodach podziemnych sieci RMWP dorzecza górnej Wisły

Statystyki

N	Ważne	20
	Braki danych	147
Średnia		1.56125
Błąd standardowy średniej		.78273
Mediana		.36250
Dominanta		.220[a]
Odchylenie standardowe		3.50050
Wariancja		12.25348
Skośność		3.491
Błąd standardowy skośności		.512
Kurtoza		12.590
Błąd standardowy kurtozy		.992
Rozstęp		14.780
Minimum		.220
Maksimum		15.000
Percentyle	16	.25072
	25	.26325
	50	.36250
	75	.79450
	84	1.51264

a. Istnieje wiele wartości modalnych. Podano wartość najmniejszą.



Cynk [mg/dm3]

Spis literatury

[1] Bednarczyk S., 1998: *Metodyka Regionalnego Monitoringu Wód Podziemnych w świetle badań na wybranym obszarze zlewni górnej Wisły*. ZHiOW AGH, praca doktorska (niepubl.)

[2] Benjamin J. R, Cornell C. A., 1977: *Rachunek prawdopodobieństwa, statystyka matematyczna i teoria decyzji dla inżynierów*. Warszawa, WNT

[3] Bieniek A., 1999: *Statystyczna kontrola jakości danych hydrogeochemicznych w monitoringu wód podziemnych*. ZHiOW AGH, praca magisterska napisana pod kierunkiem prof. Jadwigi Szczepańskiej

[4] Blaschke Z., 1995: *Ogólna charakterystyka metody „uogólnionego portretu"*. [w:] Materiały XIX sympozjum „Zastosowania metod matematycznych i informatyki w geologii", Kraków

[5] Błaszyk T., Macioszczyk A., 1993: *Klasyfikacja jakości zwykłych wód podziemnych dla potrzeb monitoringu środowiska*. Warszawa, PIOŚ, Biblioteka Monitoringu Środowiska

[6] Czermiński J. B., Iwasiewicz A., Paszek Z., Sikorski A., 1994: *Metody statystyczne dla chemików*. Warszawa, PWN

[7] Doerffel K, 1989: *Statystyka dla chemików analityków*. Warszawa, WNT

[8] Dynowska J., Maciejewski M. [Ed.], 1991: *Dorzecze górnej Wisły*. t. I, II. Warszawa-Kraków, PWN

[9] Englund E., Sparks A., 1991: *Geo-Eas 1.2.1. Geostatistical environmental assesment software. User's guide*. U.S. Environmental Protection Agency, Las Vegas, Nevada

[10] Fleming J., Albus H., Neidhart B., Wegscheider W., 1997: *Glossary of analytical terms*. Journal for quality, comparability and reliability in chemical measurement, no 2/1/97 „Terminology and definitions". Berlin, Heidelberg, Springer-Verlag

[11] Górniak J., Wachnicki J., 2000: *Pierwsze kroki w analizie danych*. SPSS Polska

[12] Gruszczyński S., 2000: *Symulacja skutków przekształceń gleb na terenach górniczych za pomocą klasyfikatorów neuronowych*. Kraków, Wydawnictwa AGH

[13] Helsel D. R., Hirsch R M., 1992: *Statistical methods in water resources*. Amsterdam / London / New York / Tokyo, ELSEVIER

[14] Hippe Z., 2000: *Data Mining and Knowledge Discovery in Chemistry: Possibilities and Limitations*. maszynopis

[15] Hordejuk T., 1993: *Krajowy monitoring wód podziemnych — organizacja, główne wyniki prac i badań*. [w:] „Biologia i monitoring wód podziemnych", Kowalczyk E., Szczepański A. [Ed.]. Częstochowa, 16–17.11.1993

[16] Hordejuk T., Gawin A., 1994: *Wyniki monitoringu jakości zwykłych wód podziemnych w latach 1991–1993 (sieć krajowa)*. Warszawa, PIOŚ, Biblioteka Monitoringu Środowiska

[17] Huber L., 1997: *Dobra praktyka laboratoryjna w analizie instrumentalnej*, Biblioteka Monitoringu Środowiska, PIOŚ, Warszawa

[18] Kalabiński J., Mastej W., 1995: *Komputerowy pakiet obliczeniowy „Metody Rozpoznawania Obrazów"*. [w:] Materiały XIX sympozjum „Zastosowania metod matematycznych i informatyki w geologii", Kraków

[19] Kania J., 2000: *Wpływ likwidacji kopalń odkrywkowych siarki na zmiany stosunków wodnych w ich otoczeniu*. ZHiOW AGH, praca doktorska (niepubl.)

[20] Kazimierski B., Sadurski A. [Ed.], 1999: *Monitoring osłonowy ujęć wód podziemnych. Metody badań*. Warszawa, PIG

[21] Kleczkowski A. S. [Ed.], 1990: *Mapa obszarów głównych zbiorników wód podziemnych (GZWP) w Polsce wymagających szczególnej ochrony*. Skala 1 : 500 000. Kraków, Wyd. AGH

[22] Kleczkowski A. S. et al., 1991: *ZTE monitoringu jakości wód podziemnych dla dorzecza górnej Wisły. Krakowski region wodnogospodarczy (Kr)*. Kraków, Inst. HiGI, AGH (niepubl.)

[23] Kleczkowski S., [Ed.], 1997: *Słownik hydrogeologiczny*, Wyd. Trio, Warszawa

[24] Kleczkowski A. S., Myszka J., Solecki T., 1994: *Krakowskie artezyjskie zdroje wód pitnych z wapieni jury*. Kraków, AGH, OW

[25] Kmiecik E., 1995: *Statystyczna kontrola jakości w oparciu o system Statgraphics*. Kraków, Uniwersytet Jagielloński, Zakład Chemii Analitycznej Wydziału Chemii, praca magisterska napisana pod kier. prof. dra hab. A. Parczewskiego

[26] Kmiecik E., 1999: *QI Analyst Gage R&R — komputerowa analiza powtarzalności i odtwarzalności systemów pomiarowych do oznaczania składników chemicznych wód*, [w:] Współczesne problemy hydrogeologii, t. IX, red. Krajewski S., Sadurski A., Warszawa–Kielce 15–17 września 1999

[27] Kmiecik E., 1999a: *Komputerowo wspomagana analiza precyzji systemów pomiarowych do oznaczania wskaźników chemicznych wód*. prace PAN (w druku)

[28] Kmiecik E., 2000: *Prediction of long-term quality transformations of leachate from coal-mining waste dump with the use of the neural networks*. Praga 2000 (publikacja na CD-ROM)

[29] Kotlarczyk J., Mastej W., Kalabiński J., Blaschke Z., 1995: *Elementy nowej strategii rozpoznawania złóż Zn–Pb w rejonie śląsko-krakowskim za pomocą metod rozpoznawania obrazów*. [w:] Materiały XIX sympozjum „Zastosowania metod matematycznych i informatyki w geologii", Kraków

[30] Kotlarczyk J., Mastej W., Kalabiński J., 1997: *Wyniki zastosowania nowej strategii rozpoznawania złóż Zn–Pb*. [w:] Materiały XX sympozjum „Zastosowania metod matematycznych i informatyki w geologii", Kraków

[31] Kotlarczyk J., Jucha S. F., Mastej W., Namysłowska-Wilczyńska B., 1999: *Rozpoznawanie obrazów w prospekcji stref naftowych w cenomanie i malmie synklinorium Nidy*. Gospodarka Surowcami Mineralnymi 1999, t. 15, z. spec. 45–68

[32] Kropka J., Różkowski A., 1994: *Wstępne wyniki regionalnego monitoringu jakości wód triasowych zbiorników wód podziemnych*. [W:] „Zaopatrzenie w wodę miast i wsi", Sozański M. [Ed.]. Poznań

[33] Lachtermacher G., Fuller J. D.: *Backpropagation in hydrological times series forecasting*. [w:] Hipel K. W., McLeod A. I., Panu U. S., Sing V. P. [Ed.]: *Stochastic and statistical methods in hydrology and environmental engineering*. vol. 3. „Time series analysis in hydrology and environmental engineering". Dodrecht/Boston/Londyn, Kluwer Academic Publishers 1994

[34] Lula P., 2001: *Wykorzystanie sztucznej inteligencji w prognozowaniu*. Statsoft Polska sp. z o.o. (publikacja elektroniczna na stronach http:
www.statsoft.com.pl)

[35] Luszniewicz A., Słaby T., 1997: *Statystyka stosowana*. PWE, Warszawa

[36] Macioszczyk A., 1987: *Hydrogeochemia*. Warszawa, Wydawnictwa Geologiczne

[37] Macioszczyk A., 1990: *Tło i anomalie hydrogeochemiczne. Metody badania, oceny i interpretacji*. CPBP 04.10.09, z. 54 (maszynopis)

[38] Macioszczyk A., Witczak S., 1999: *Współczesne problemy hydrogeochemii*. Biuletyn PIG, nr 388, Wyd. PIG, Warszawa, s. 139–156

[39] Mcculloch W. S., Pitts W., 1943: *A logical calculus of the ideas immanent in nervous activity*. Bulletin of Mathematical Biophysics, No 5, 1943, pp. 115–133

[40] Mucha J., 1991: *Wybrane metody matematyczne w geologii górniczej*. Kraków, Wyd. AGH

[41] Nabagło I., 1994: *Zastosowanie sieci neuronowych do predykcji nieliniowych sygnałów losowych*. [w:] Materiały konferencyjne I krajowej konferencji „Sieci neuronowe i ich zastosowania". T. II. Częstochowa

[42] Nielsen D. M., 1991: *Practical handbook of ground-water monitoring*. Chelsea, Lewis Publishers

[43] Osmęda-Ernst E., Szczepańska J., Witczak S., 1995: *Praktyczna granica oznaczalności (PQL) jako kryterium jakości opróbowania w monitoringu wód podziemnych*. [w:] „Współczesne problemy hydrogeologii". Szczepańska J., Kulma R., Szczepański A. [Ed.]. t. VII. Kraków

[44] Osmęda-Ernst E., Bobrowski A., Rzepecki T., Gajewska I., Knap W., 1996: *Znaczenie granic wykrywalności i oznaczalności w analizie mikroskładników wód podziemnych*. Materiały konferencyjne VII konferencji „Analityka w służbie geologii i ochrony środowiska". Szelment, 17–21 czerwca 1996

[45] Pasiut D., 2000: *Statystyczna kontrola jakości...* ZHiOW AGH, praca magisterska napisana pod kierunkiem prof. Jadwigi Szczepańskiej

[46] Petridis V., Kehagias A., 1998: *Predictive modular neural networks. Applications to Time Series*. Boston/Dodrecht/Londyn, Kluwer Academic Publishers

[47] Prażak J., Janecka-Styrcz K., Kowalczewska G., Paciura W., 1996: *Raport o jakości zwykłych wód podziemnych województwa kieleckiego na podstawie badań monitoringowych wykonanych w latach 1991–1995*. Kielce, PIOŚ, Biblioteka Monitoringu Środowiska

[48] Ramsey M. H., 1992: *Sampling and Analytical Quality Control (SAX) for improved error estimation in the measurement of Pb in the environment using robust analysis of variance*. Applied Geochemistry. Suppl. Issue no 2

[49] Ramsey M. H., Thompson M., Hale M., 1992: *Objective evaluation of precision requirements for geochemical analysis using robust analysis of variance*. J. Geochem. Explor, 44

[50] Różkowski A. et al., 1991: *ZTE monitoringu jakości wód podziemnych dla dorzecza górnej Wisły. Katowicki region wodnogospodarczy (Ka)*. Sosnowiec, INTERGEO (niepubl.)

[51] Siwek P., 1999: *Chemizm i jakość wód podziemnych serii węglanowej zbiornika triasu gliwickiego w świetle monitoringu regionalnego*. Uniwersytet Śląski, Wydz. Nauk o Ziemi, praca doktorska (niepubl.)

[52] Staniewicz-Dubois H., 1991: *Wskazówki metodyczne dotyczące tworzenia regionalnych i lokalnych monitoringów wód podziemnych*, wyd. I. PIOŚ. Biblioteka Moniotirngu Środowiska. Warszawa

[53] Staniewicz-Dubois H., 1995: *Wskazówki metodyczne dotyczące tworzenia regionalnych i lokalnych monitoringów wód podziemnych*, wyd. II zmienione. PIOŚ. Biblioteka Monitoringu Środowiska. Warszawa

[54] Szczepańska J., Witczak S., Postawa A., 1996: *Zastosowanie analizy wariancji (ANOVA) do oceny precyzji wyników badań hydrogeochemicznych*. [w:] „Problemy hydrogeologiczne południowo-zachodniej Polski", Ciężkowski W. [Ed.]. Wrocław

[55] Szczepańska J., Witczak S., Postawa A., 1996a: *Zastosowanie analizy wariancji (ANOVA) do oceny jako×ci badań w monitoringu wód podziemnych*. [w:] „Technika i technologia w ochronie środowiska". I Forum Inżynierii Ekologicznej. Wiatr I. [Ed.]. Lublin-Nałęczów

[56] Szczepańska J., Witczak S., Postawa A., Knap W., 1997: *Zapewnienie jakości/kontrola jakości QA/QC badań hydrogeochemicznych w monitoringu wód podziemnych*. [w:] „Współczesne problemy hydrogeologii". Górski J., Liszkowska E. [Ed.]. t. VIII. Poznań

[57] Szczepańska J., Kmiecik E., 1998: *Statystyczna kontrola jakości danych w monitoringu wód podziemnych*. Kraków, Wydawnictwa AGH

[58] Szczepańska J., Kmiecik E., 1998a: *Ocena precyzji oznaczeń cynku w próbkach wody w warunkach powtarzalności i odtwarzalności*. Prace Naukowe Uniwersytetu Śląskiego w Katowicach Nr 1718. Wydawnictwa Uniwersytetu Śląskiego, Katowice

[59] Szczepańska J., Kmiecik E., 1998b: *Ocena precyzji oznaczeń cynku metodą absorpcyjnej spektrometrii atomowej AAS w wodzie podziemnej ze zdroju Królewskiego w Krakowie*. [w:] II Forum Inżynierii Ekologicznej. Wiatr I. [Ed.]. Lublin–Nałęczów

[60] Szczepańska J., Kmiecik E., 2000: *Prognozowanie zmian jakości wód w układzie czasowym z wykorzystaniem sieci neuronowych*. [w:] Lublin–Nałęczów

[61] Szczepańska J., Kmiecik E., 2001: *Wykorzystanie sieci neuronowych do oceny czasu oddziaływania składowiska odpadów górniczych na środowisko wodne*. (w druku)

[62] Świercz M., 1994: *Application of neural networks to demand prediction in water distribution networks*. [w:] Materiały konferencyjne I krajowej konferencji „Sieci neuronowe i ich zastosowania". T. II. Częstochowa

[63] Tadeusiewicz R., 1993: *Sieci neuronowe*. Akademicka Oficyna Wydawnicza RM, Warszawa

[64] Tadeusiewicz R., Mikrut R., 1994: *Sieci neuronowe rozpoznające obrazy*. [w:] I Kraj. Konferencja „Sieci neuronowe i ich zastosowania" 12–15.04.1994

[65] Tadeusiewicz R., 1999: *Wprowadzenie do praktyki stosowania sieci neuronowych*. [w:] Sieci Neuronowe. Materiały na seminarium organizowane przez Statsoft Polska sp. z o.o. 14.10.1999 w Warszawie

[66] Tadeusiewicz R., 2001: *Wprowadzenie do praktyki stosowania sieci neuronowych*. Statsoft Polska sp. z o.o. 2001 (publikacja elektroniczna na stronach http://www.statsoft.com.pl)

[67] Thompson M., Howarth R. J., 1976: *Duplicate analysis in geochemical practice*. P. 2. Theoretical approach and estimation of analytical reproducibility. Analyst, Sept. 1976, vol. 101

[68] Waksmundzki T., 1995: *Algorytm LI — „najmniejszego przedziału" — próba zastosowania przy rozpoznawaniu złóż*. [w:] Materiały XIX sympozjum „Zastosowania metod matematycznych i informatyki w geologii", Kraków

[69] Wiatr I., 1998: *Wstęp* do II Forum Inżynierii Ekologicznej „Monitoring Środowiska" [red.] Wiatr I., Marczak H., Nałęczów

[70] Witczak S. et al., 1993: *The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB)*. PHARE Regional Environmental Sector Programme 1991. Techn. Serv. Contract no P-UV/2. Kraków, AGH

[71] Witczak S. et al., 1993a: *The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB)*. Report for the first quarter of 1993. Kraków, AGH

[72] Witczak S. et al., 1993b: *The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB)*. Report for the second quarter of 1993. Kraków, AGH

[73] Witczak S. et al., 1993c: *The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB)*. Report for the third quarter of 1993 + Annex: Documentation of GQM points for the area of Kraków and Katowice Regional Council for Water Management. Kraków, AGH

[74] Witczak S. et al., 1993d: *The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB)*. Report for the fourth quarter of 1993 + Annex: Technical adaptation of selected wells and spring for the area of Kraków and Katowice Regional Council for Water Management, Kraków, AGH

[75] Witczak S., Adamczyk A., 1994: *Katalog wybranych fizycznych i chemicznych wskaźników zanieczyszczeń wód podziemnych i metod ich oznaczania*. t. I. Warszawa, PIOŚ, Biblioteka Monitoringu Środowiska

[76] Witczak S. et al., 1994a: *The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB)*. Final Report (Text) + Figures + Tables + Appendices 1, 2. Kraków, AGH

[77] Witczak S. et al., 1994b: *Monitoring jakości wód podziemnych w dorzeczu górnej Wisły — obszar wschodni (badania w zakresie kontroli QA/QC)*. Kraków, AGH

[78] Witczak S., Adamczyk A., 1995: *Katalog wybranych fizycznych i chemicznych wskaźników wód podziemnych i metod ich oznaczania*. t. II. Warszawa, PIOŚ, Biblioteka Monitoringu Środowiska

[79] Witkowski A., 1997: *Monitoring jakości zwykłych wód podziemnych w obszarze działania Regionalnego Zarządu Gospodarki Wodnej w Katowicach. Raport z badań wykonanych w latach 1993–1996*. Katowice, RZGW Katowice, Wydział Nauk o Ziemi UŚ

[80] Zamorska J., 1999: *Prognozowanie wybranych właściwości wody w środowisku naturalnym metodą rozpoznawania obrazów*. Rozprawa doktorska (niepubl.), ZhiOW AGH, Kraków

[81] Zhang S. P., Watanabe H., Yamada R., 1994: *Prediction of daily water demands by neural networks*. [w:] Hipel K. W., McLeod A. I., Panu U. S., Sing V. P. [Ed.]: *Stochastic and statistical methods in hydrology and environmental engineering*. vol. 3. „Time series analysis in hydrology and environmental engineering". Dodrecht/Boston/Londyn, Kluwer Academic Publishers

[82] Zhu Mu-Lan, Fujita M., Hashimoto N., 1994: *Application of neural networks to runoff prediction*. [w:] Hipel K. W., McLeod A. I., Panu U. S., Sing V. P. [Ed.]: *Stochastic and statistical methods in hydrology and environmental engineering*. vol. 3. „Time series analysis in hydrology and environmental engineering". Dodrecht/Boston/Londyn, Kluwer Academic Publishers

[83] Dyrektywa Unii Europejskiej 98/83/EC — Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption

[84] Dyrektywa Unii Europejskiej 2000/60/EC — Directive of the european parliament and of the council of 23 October 2000 establishing a framework for community action in the field of water policy

[85] Dz. U. nr 82, poz. 937 — Rozporządzenie Ministra Zdrowia z dnia 4 września 2000 roku, w sprawie warunków, jakim powinna odpowiadać woda do picia i na potrzeby gospodarcze, woda w kąpieliskach, oraz zasad sprawowania kontroli jakości wody przez organy Inspekcji Sanitarnej

[86] GRID, 1993: Andrzejewski R., Baranowski M. [red.] Stan środowiska w Polsce. Warszawa, Centrum Informacji o Środowisku

[87] *Międzynarodowy słownik podstawowych i ogólnych terminów metrologii (International Vocabulary of Basic and General Terms in Metrology)*. Główny Urząd Miar, 1996

[88] Millipore, 1997 - Laboratory Cataloque. Millipore, o. w Warszawie, ul. Jasnodworska 7

[89] Monitor Polski Nr 6 z 1991 r. Zarządzenie MOŚZNiL z 1 II 1991 roku w sprawie utworzenia regionalnych zarządów gospodarki wodnej

[90] *Prawo ochrony środowiska Wspólnoty Europejskiej*. t. VII. Woda. Warszawa, Ministerstwo Ochrony Środowiska Zasobów Naturalnych i Leśnictwa, 1996

[91] SPSS, 1997, 1999 — dokumentacja do programu Neural Connection, v. 2.0, v. 2.1

[92] SPSS Inc., 1997a: *Dokumentacja do programów SPSS v. 7.5, QI Analyst 3.5 DB*

[93] SPSS Inc., 2000: *Dokumentacja do programu SPSS v. 10.0 PL*

[94] USP XXIII (United States Pharmacopeia): *Validation of compendial methods*. pp. 1982–1984

[95] WHO, 1998: *Wytyczne WHO dotyczące jakości wody do picia*. tom. I. Zalecenia. Wyd. II. Polskie Zrzeszenie Inżynierów i Techników Sanitarnych

Skorowidz