BIULETYN PAŃSTWOWEGO INSTYTUTU GEOLOGICZNEGO 412: 5–70, 2004 R.

Ewa KMIECIK[1]

PROGNOZOWANIE ZMIAN JAKOŚCI WÓD PODZIEMNYCH W UKŁADZIE PRZESTRZENNYM Z WYKORZYSTANIEM SIECI NEURONOWYCH

(z 37 fig.)

SPATIAL PREDICTIONS OF GROUNDWATER QUALITY CHANGES USING NEURAL NETWORKS

(with 37 Figs.)

Abstract. This paper presents using neural networks in spatial prediction of groundwater quality changes on the base of existing database. This database consists of results of regional groundwater quality monitoring of the upper Vistula river basin carried out in 1993–1994 (Witczak *et al.*, 1994a, b).

Data (the results of field and laboratory determinations of physicochemical indicators of groundwater quality) was verified using quality control parameters and statistical analysis.

On the verified database were conducted predictive trials to provide values of physicochemical indicators for the monitoring sites with known coordinates and monitoring site classification (on the base of physicochemical indicators values) to the area of known type of land-use.

The results of such a study show that neural networks can be succesfully used for spatial prediction of changes in groundwater quality. The condition for reliability of the prognoses is verification of input data loaded to the model.

Keywords: groundwater quality, monitoring networks, hydrogeochemical data, neural networks, prediction, classification.

Abstrakt. Zastosowanie 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 i in., 1994a, b).

Wyniki oznaczeń terenowych i laboratoryjnych (55) wskaźników fizykochemicznych (nieorganicznych i organicznych) wód poddano weryfikacji z zastosowaniem parametrów kontroli jakości oraz statystycznej analizy rozkładu tych wskaźników.

[1]Akademia Górniczo-Hutnicza, Zakład Hydrogeologii i Ochrony Wód, al. Mickiewicza 30, 30-059 Kraków

Na zweryfikowanej bazie danych przeprowadzono próby predykcji wartości wskaźników fizykochemicznych wód dla punktu monitoringowego o określonych współrzędnych oraz klasyfikacji punktu monitoringowego (na podstawie wyników oznaczeń wskaźników fizykochemicznych) do obszaru o określonym użytkowaniu terenu.

Uzyskane wyniki badań wskazują, że sieci neuronowe można z powodzeniem wykorzystać do prognozowania zmian jakości wód w układzie przestrzennym. Warunkiem jednak, by uzyskiwane prognozy cechowały się wysokim stopniem wiarygodności, jest konieczność weryfikacji danych wejściowych wprowadzanych do modelu.

Słowa kluczowe: jakość wód podziemnych, sieci monitoringowe, dane hydrogeochemiczne, sieci neuronowe, predykcja, klasyfikacja

WSTĘP

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... (Słownik hydrogeologiczny, 2002, s. 134–135).

Coraz większe znaczenie wód podziemnych, bardzo często jedynego źródła wód pitnych o dobrej jakości, wymusiło podjęcie w Polsce badań jakości tych wód w trzech rodzajach sieci monitoringowych: krajowej, regionalnych i lokalnych (Hordejuk, 1993; Hordejuk, Gawin, 1994; Kropka, Różkowski, 1994; Witczak i in., 1994a, b; Prażak i in., 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 cechować się wysokim stopniem pewności i stąd wynika konieczność ciągłej kontroli ich jakości, tak w odniesieniu do badań laboratoryjnych (norma PN/EN ISO 17025), jak i procesu opróbowania.

Celem niniejszej pracy jest udowodnienie iż:

- poprzez analizę rozkładu badanych wskaźników fizykochemicznych 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, wykorzystując wyniki badań (wartości wskaźników fizykochemicznych wód podziemnych) przeprowadzonych w punktach sąsiednich (zagadnienia **predykcji**);
- za pomocą sieci neuronowych, na podstawie wartości wskaźników fizykochemicznych wód w danym punkcie monitoringowym można dokonać klasyfikacji przynależności tego punktu do obszaru o określonym zagospodarowaniu terenu (zagadnienia **klasyfikacji**).

Eksperymenty komputerowe w zakresie weryfikacji danych oraz prognozowania zmian jakości wód podziemnych w układzie przestrzennym zostały wykonane na wynikach badań przeprowadzonych w latach 1993–1994, w sieci regionalnego monitoringu jakości wód podziemnych dorzecza górnej Wisły (Witczak i in., 1993a, b, c, d, e; Witczak i in., 1994a, b; Witkowski, 1997). Weryfikacja punktów monitoringowych odbyła się na etapie projektowania sieci poprzez wykorzystanie kartowania sozologicznego, z uwzględnieniem zagospodarowania terenu i warunków hydrogeologicznych.

Dane hydrogeochemiczne (oznaczenia 55 wskaźników fizykochemicznych wód) zostały poddane potrójnej weryfikacji: porównano granice oznaczalności badanych wskaźników (laboratoryjne DL i praktyczne PDL), oszacowano udział wariancji technicznej σ^2_{tech} w wariancji całkowitej σ^2_{tot} na podstawie wyników badań próbek dublowanych, wykorzystując klasyczną analizę wariancji ANOVA oraz elastyczne postępowanie statystyczne (*robust statistics*) a następnie dokonano statystycznej analizy rozkładu tych wskaźników. Z dalszej analizy zostały wyłączone obserwacje anomalne, obarczone błędami grubymi, oznaczenia wskaźników fizykochemicznych cechujące się niską precyzją, i oznaczenia tych wskaźników, w zbiorze których ponad 20% stanowiły wyniki poniżej granicy oznaczalności DL.

Tak zweryfikowaną bazę danych dla sieci RMWP dorzecza górnej Wisły (16 wskaźników fizykochemicznych spośród 55 analizowanych) wykorzystano do prognozowania zmian jakości wód w układzie przestrzennym za pomocą sieci neuronowych.

Przygotowano trzy warianty danych zweryfikowanych, różniące się liczbą zmiennych (wskaźników fizykochemicznych) i obserwacji (punktów RMWP):

- zbiór zawierający wszystkie zweryfikowane wskaźniki fizykochemiczne (16) i punkty monitoringowe o klasach zagrożenia wód AB, C, D (wyodrębionych na podstawie czasu migracji wody z powierzchni terenu do monitorowanej warstwy wodonośnej; 167 punktów RMWP);
- zbiór zawierający wszystkie zweryfikowane wskaźniki fizykochemiczne (16), a 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 predykcji i klasyfikacji jakości wód podziemnych w układzie przestrzennym.

Różne warianty danych wejściowych umożliwiły ocenę wpływu na jakość uzyskiwanych prognoz liczby zweryfikowanych wskaźników fizykochemicznych 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.

Modele sieci neuronowych budowano i testowano w programie Neural Connection (SPSS, 1997, 1999), udostępnionym nieodpłatnie dla celów pisania rozprawy doktorskiej (Kmiecik, 2001) oraz niniejszej pracy, dzięki uprzejmości prezesa firmy SPSS Polska sp. z o.o. Rozdział zawierający teorię z zakresu sieci neuronowych oparty jest w głównej mierze na dokumentacji dotyczącej wersji 2.1 programu (SPSS, 1997).

Najlepsze wyniki prognoz wartości wskaźników fizykochemicznych wód na podstawie współrzędnych punktu monitoringowego — najmniejsze błędy względne prognoz — uzyskano dla sieci RBF.

Pliki z analizowanymi danymi, pliki z modelami budowanych sieci neuronowych oraz pliki wynikowe i raporty z przeprowadzonych analiz w formatach programów SPSS i Neural Connection znajdują się w publikacji elektronicznej (Kmiecik, 2001) na stronie internetowej http://galaxy.agh.edu.pl/~ek. Dostęp do nich umożliwi zainteresowanemu Czytelnikowi samodzielne przeprowadzenie przedstawionych w niniejszej pracy analiz.

REGIONALNY MONITORING JAKOŚCI WÓD PODZIEMNYCH (RMWP) DORZECZA GÓRNEJ WISŁY

Do prognozowania zmian jakości wód podziemnych w układzie przestrzennym (poziomym) wykorzystano wyniki badań przeprowadzonych w latach 1993–1994, w sieci Regionalnego Monitoringu Jakości Wód Podziemnych (RMWP) dorzecza górnej Wisły. Sieć ta składa się ze 172 punktów RMWP, z czego w obszarze Regionalnego Zarządu Gospodarki Wodnej (RZGW) Kraków znajduje się 117, zaś w obszarze RZGW Katowice — 55 punktów (tab. 1).

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

CHARAKTERYSTYKA SIECI MONITORINGOWEJ

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. Program pilotowy objął obszar dwóch spośród siedmiu utworzonych wówczas (Monitor Polski Nr 6 z 1991 r., poz. 38) Regionalnych Zarządów Gospodarki Wodnej — RZGW Katowice i RZGW Kraków (fig. 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: Wojewódzkiego Inspektoratu Ochrony Środowiska (WIOŚ) w Tarnowie (dla obszaru RZGW Kraków) i Ośrodka Badań i Kontroli Środowiska (OBiKŚ) w Katowicach (dla obszaru RZGW Katowice).

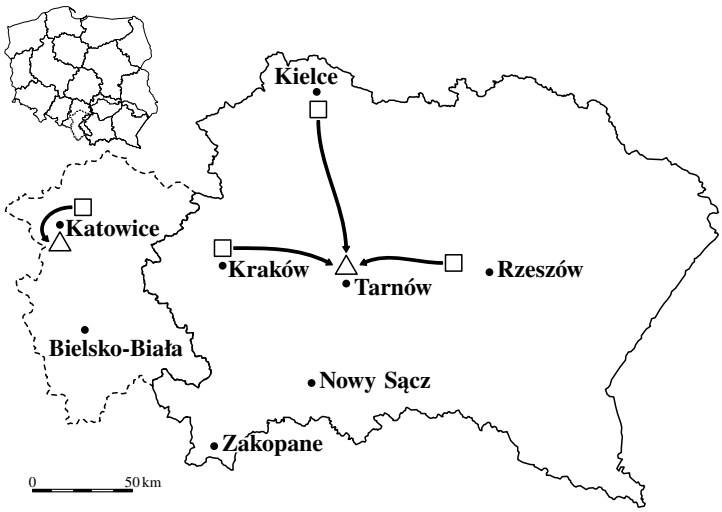


Fig. 1. Lokalizacja regionalnego monitoringu jakości wód podziemnych (RMWP) w zlewni górnej Wisły
- - - granice RZGW Katowice, — granice RZGW Kraków (wg Monitora Polskiego Nr 6 z 1991 r., poz. 38); □ — lokalizacja laboratoriów polowych opróbowania wód podziemnych; △ — laboratoria środowiskowe badania jakości wód podziemnych w Katowicach i Tarnowie; → — kierunki transportu próbek RMWP

Location of regional groundwater quality monitoring (RGQM) in the upper Vistula river basin
- - - boundaries of Regional Council for Water Management Katowice, — boundaries of Regional Council for Water Management Kraków (according to Monitor Polski Nr 6 z 1991 r., poz. 38)); □ — field laboratories for groundwater sampling; △ — environmental laboratories for groundwater analysis in Katowice and Tarnów; → — directions of GQM samples transportation

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 Główne Zbiorniki Wód Podziemnych (GZWP), z czego 14 występuje w obszarze RZGW Katowice, 32 GZWP w obszarze RZGW Kraków, zaś 7 GZWP przynależy do obu RZGW. Wody podziemne występują w utworach czwartorzędu, trzeciorzędu, kredy, jury, triasu, karbonu i dewonu.

Tabela 1

Charakterystyka ilościowa punktów sieci regionalnego monitoringu jakości wód podziemnych (RMWP) dorzecza górnej Wisły (wg Witczak i in., 1994a, b)

Quantitative characteristics of the sites of regional groundwater quality monitoring (RGQM) of the upper Vistula river basin (after Witczak *et al*. 1994a, b)

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 RMWP 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 (tab. 1). Szczegółowa charakterystyka punktów tworzących sieć RMWP została przedstawiona w wielu pracach (Witczak i in., 1994a, b; Witkowski, 1997; Bednarczyk, 1998; Szczepańska, Kmiecik, 1998; Siwek 1999). 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.

Wszystkie punkty tworzące sieć RMWP spełniały następujące kryteria (Witczak i in., 1994a, b):

- 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.

Przy lokalizacji punktów RMWP uwzględniono ich reprezentatywność dla określonego typu antropopresji związanej z użytkowaniem terenu (fig. 2). 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 w zlewni górnej Wisły (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.

Wartości te są zbliżone do form użytkowania terenu 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 obszarowych zanieczyszczeń atmosfery.

W monitoringu jakości wód podziemnych dorzecza górnej Wisły 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 (Witczak i in., 1994a, b) (fig. 3):

- 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.

Wśród punktów tworzących sieć RMWP, zgodnie z przyjętymi założeniami, dominuje klasa AB — 90,1%. 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.

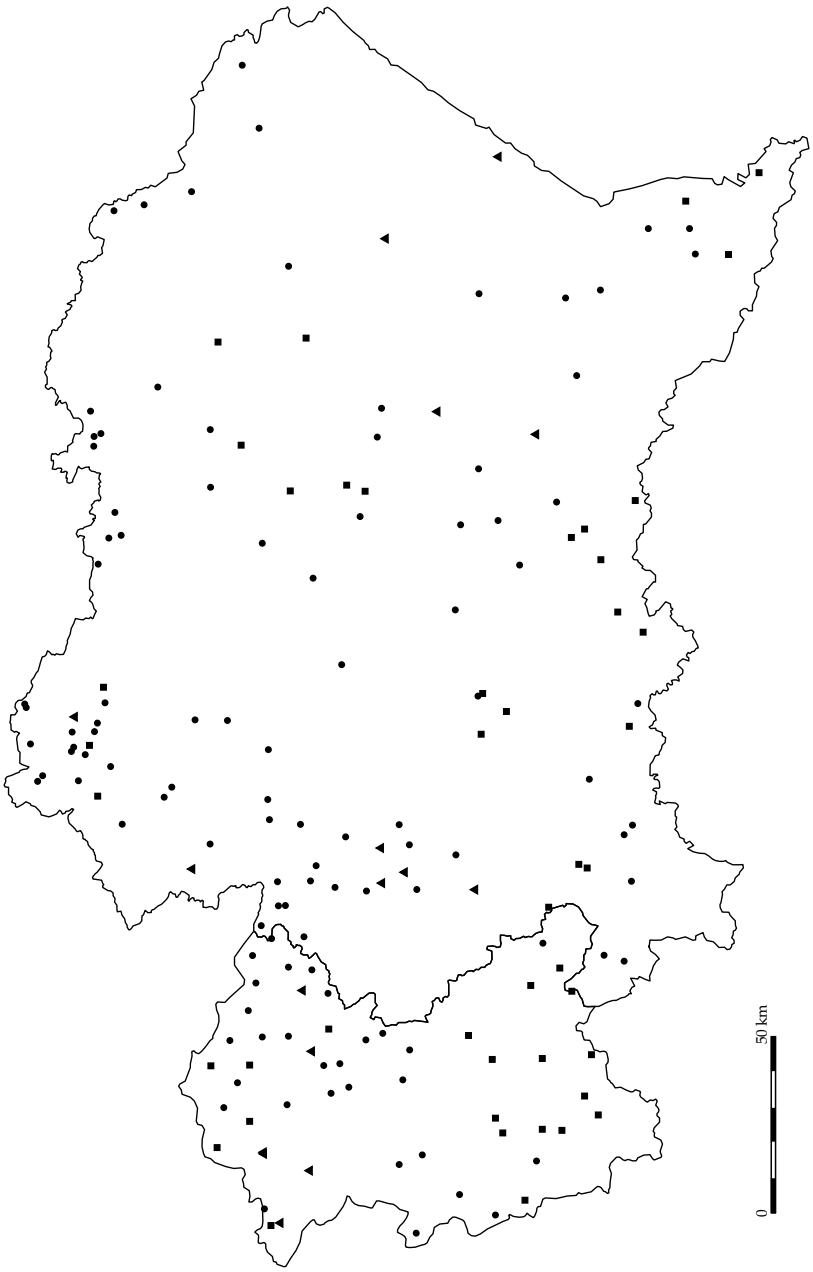


Fig. 2. Regionalny monitoring jakości wód podziemnych (RMWP) w zlewni górnej Wisły, przynależność punktów monitoringowych do różnych obszarów zagospodarowania terenu

• — R rolnicze; ■ — L leśne; ▲ — O-P osiedlowo-przemysłowe

Regional groundwater quality monitoring (RGQM) of the upper Vistula river basin, location of monitoring sites in the areas of different land-use

• — R agricultural; ■ — L forest; ▲ — O-P settlement-industrial

Fig. 3. Regionalny monitoring jakości wód podziemnych (RMWP) w zlewni górnej Wisły, przynależność punktów monitoringowych do różnych klas zagrożenia wód

• — klasa AB, wody zagrożone; ■ — klasa C, wody słabo zagrożone; ▲ — klasa D, wody praktycznie niezagrożone

Regional groundwater quality monitoring (RGQM) of the upper Vistula river basin, location of monitoring sites in the areas of different groundwater endangerment class

• — class AB, very high to medium; ■ — class C, low; ▲ — class D, very low

SPRZĘT, METODYKA ORAZ ZAKRES OZNACZEŃ ANALITYCZNYCH

Opróbowanie sieci RMWP 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 (fig. 1). Próbki filtrowano w terenie, bezpośrednio przy ich poborze z punktów przez filtr membranowy 0,45 μm. Filtrację prowadzono w systemie *on line*, bez styku wody z powietrzem.

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 analiz realizowano zgodnie z ZTE (Kleczkowski i in., 1991; Różkowski i in., 1991) oraz wytycznymi Państwowej Inspekcji Ochrony Środowiska (PIOŚ) (Staniewicz-Dubois, 1991, 1995; Błaszyk, Macioszczyk, 1993). Analiza obejmowała składniki nietrwałe, oznaczane zgodnie z wytycznymi PIOŚ (Witczak, Adamczyk, 1994, 1995) obligatoryjnie w terenie przy poborze próbki wody (10 wskaźników — temperatura, przewodność elektrolityczna właściwa, pH, potencjał utleniająco-redukcyjny Eh, mętność, osad w leju Imhoffa, barwa, zapach, zasadowość ogólna i mineralna, kwasowość ogólna) oraz składniki nieorganiczne i organiczne, oznaczane w laboratorium.

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

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 i in., 1993; Witczak, Adamczyk 1994; Szczepańska, Kmiecik, 1998).

Program terenowy QA/QC w dorzeczu górnej Wisły 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 (w tym 4,7% stanowiły próbki zerowe — wykorzystane do wyznaczenia praktycznej granicy oznaczalności PDL; 13,7% próbki dublowane — służące do oceny precyzji oznaczeń; 4,2% próbki znaczone).

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 i in., 1994a, b) oraz pracach: Witkowski (1997) i Bednarczyk (1998). W niniejszej pracy zostały one wykorzystane do weryfikacji danych wejściowych wprowadzanych do modelu sieci neuronowej.

Granica oznaczalności DL i praktyczna granica oznaczalności PDL

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 (granica oznaczalności DL) 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.

Obliczone wartości PDL zestawiono w tabeli 2 wraz z granicami oznaczalności DL deklarowanymi przez laboratoria i maksymalnymi dopuszczalnymi stężeniami MPL analizowanych wskaźników w wodach pitnych (zgodnie z rozporządzeniem Ministra Zdrowia Dz.U. Nr 203, poz. 1718 z 19 listopada 2002 roku oraz Dyrektywą Unii Europejskiej 98/83/EC z 3 listopada 1998 roku).

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 rutynowego opróbowania, 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 (czyli PDL/DL = 1).

Z tabeli 2 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 w procesie opróbowania, przechowywania i transportu (Witczak i in., 1994a, b). 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 tej granicy.

Tabela 2

Granice oznaczalności DL i praktyczne granice oznaczalności PDL (wg Witczak i in., 1994a, b) oraz maksymalne dopuszczalne stężenia MPL wybranych wskaźników w wodach pitnych wg polskich przepisów (Dz.U. Nr 203 z 2002 r., poz. 1718) i wytycznych Unii Europejskiej (Dyrektywa Unii Europejskiej 98/83/EC)

Limits of detection DL, practical determination limits PDL (according to Witczak *et al.*, 1994a, b) and maximum permissible levels MPL of selected indicators according to Polish legislation (Dz.U. Nr 203 z 2002 r., poz. 1718) and European Union Directive (98/83/EC)

Lp.	Analizowana zmienna	Jednostka	DL		PDL		PDL/DL		MPL
			Tarnów	Katowice	Tarnów	Katowice	Tarnów	Katowice	
1.	Suma subst. rozp.	mg/dm^3	1	5	4	5	4	1	—
2.	Zasadowość ogólna	mval/dm^3	0,05	0,1	0,15	0,1	3	1	—
3.	Twardość ogólna	mval/dm^3	2	4	2	4	1	1	60–500
4.	Potas	mg/dm^3	0,01	0,2	1,5	0,7	150	3,5	—
5.	Sód	mg/dm^3	0,1	0,2	0,1	0,8	1	4	200
6.	Magnez	mg/dm^3	0,1	5	0,7	5	7	1	50
7.	Wapń	mg/dm^3	0,1	0,5	4,8	5	48	10	—
8.	Azot amonowy	mg/dm^3	0,04	0,01	0,1	0,15	2,5	15	0,5
9.	Glin	mg/dm^3	0,015	0,01	0,08	0,15	5,3	15	0,2
10.	Żelazo ogólne	mg/dm^3	0,03	0,01	2,2	0,27	73,3	27	0,2
11.	Mangan	mg/dm^3	0,01	0,01	0,03	0,085	3	8,5	0,05
12.	Azot azotynowy	mg/dm^3	0,001	0,001	0,003	0,007	3	7	0,1
13.	Azot azotanowy	mg/dm^3	0,1	0,1	0,2	—	2	—	50
14.	Chlorki	mg/dm^3	5	0,5	5	5	1	10	250
15.	Siarczany	mg/dm^3	10	10	10	10	1	1	250
16.	Fosforany rozpuszczone	mg/dm^3	0,05	0,05	0,05	0,05	1	1	5*
17.	Krzemionka zdysocjowana	mg/dm^3	0,7	0,5	0,7	0,5	1	1	—
18.	Fluorki	mg/dm^3	0,1	0,01	0,1	0,01	1	1	1,5
19.	Bor	mg/dm^3	0,005	—	0,23	—	46	—	1
20.	Chrom ogólny	mg/dm^3	0,003	0,01	0,01	0,01	3,3	1	0,05
21.	Cynk	mg/dm^3	0,01	0,01	0,075	0,11	7,5	11	3
22.	Kadm	mg/dm^3	0,001	0,001	0,001	0,001	1	1	0,003
23.	Miedź	mg/dm^3	0,002	0,01	0,01	0,01	5	1	1
24.	Nikiel	mg/dm^3	0,001	0,01	0,02	0,01	20	1	0,02
25.	Ołów	mg/dm^3	0,001	0,005	0,0068	0,016	6,8	3,2	0,01
26.	Rtęć	mg/dm^3	0,0002	0,0002	0,0037	0,0002	18,5	1	0,001
27.	Współcz. absorpcji UV (A254)		0,005	0,005	0,032	—	6,4	—	—
28.	Rozp. węgiel organiczny	mg/dm^3	0,2	0,5	2,33	—	11,7	—	—
29.	Utlenialność ChZT-Mn	mg/dm^3	0,5	0,5	1,12	—	2,2	—	5
30.	Azot organiczny Kjeldahla	mg/dm^3	0,5	0,5	0,5	0,5	1	1	—
31.	Fenole lotne	mg/dm^3	0,001	0,01	0,001	0,01	1	1	0,0005
32.	Substancje ropopochodne	mg/dm^3	0,05	0,01	0,45	—	9	—	—
33.	Chloroform	mg/dm^3	0,00001	0,0001	0,399	—	39900	—	0,03
34.	Subst. pow.-czynne anionowe	mg/dm^3	0,0001	0,0001	0,0001	0,0001	1	1	0,2
35.	Czterochloroetylen	mg/dm^3	0,000005	0,000005	0,0056	—	1120	—	0,01
36.	Trójchloroetylen	mg/dm^3	0,00003	0,00001	0,019	0,0028	633,3	280	0,05
37.	DDT	mg/dm^3	0,00002	0,000005	0,002	0,000005	100	1	—
38.	DDE	mg/dm^3	0,000008	0,000005	0,0004	0,000005	50	1	—
39.	DDD	mg/dm^3	0,000008	0,000005	0,003	0,000006	375	1,2	—
40.	Gamma-HCH	mg/dm^3	0,000008	0,000003	0,000009	0,000004	1,1	1,3	—
41.	Metoksychlor	mg/dm^3	0,00005	0,000008	0,001	0,000008	20	1	—
42.	Benzo-a-piren	mg/dm^3	—	—	—	—	—	—	—
43.	Suma 6WWA	mg/dm^3	—	—	—	—	—	—	—

*fosfor jako P_2O_5; długa kreska (—) oznacza brak danych

Precyzja 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. Do opisu precyzji można więc stosować miary rozrzutu: rozstęp, odchylenie standardowe, wariancję, współczynnik zmienności.

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 i in., 1992). Dotychczasowe doświadczenia związane z realizacją terenowego programu kontroli QA/QC (Witczak i in., 1994a, b; Osmęda-Ernst i in., 1995, 1996; Szczepańska i in., 1996a, b, 1997) wskazują, że jest to najtańsza, najszybsza i najlepsza metoda oszacowania błędów losowych powstających w procesach opróbowania i/lub analityki.

Terenowy program kontroli jakości QA/QC przeprowadzony w pierwszej serii opróbowania sieci RMWP dorzecza górnej Wisły umożliwił obliczenie za pomocą programu komputerowego ROB 2, wariancji technicznej σ^2_{tech} (uwzględniającej łączny wpływ błędów opróbowania σ^2_{sampl} i analityki σ^2_{anal}) metodą klasyczną i metodą *robust statistics* (elastycznego postępowania statystycznego, bez odrzucania błędów grubych):

$$\sigma^2_{tot} = \sigma^2_{geochem} + \sigma^2_{sampl} + \sigma^2_{anal} = \sigma^2_{geochem} + \sigma^2_{tech}$$

$$\sigma^2_{tech} = \sigma^2_{sampl} + \sigma^2_{anal}$$

(σ^2_{tot} — wariancja całkowita; $\sigma^2_{geochem}$ — wariancja hydrogeochemiczna).

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 (< DL = DL; Helsel, Hirsch, 1992). 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 3 i 4 zestawiono procentowe udziały wariancji technicznej σ^2_{tech} w wariancji całkowitej σ^2_{tot}, obliczone metodą klasycznej analizy wariancji ANOVA i elastycznego postępowania statystycznego ROBUST. 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. 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. 3) 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.

Tabela 3

RMWP dorzecza górnej Wisły — pierwsza seria opróbowania. Wariancja techniczna σ^2_{tech} obliczona metodą klasycznej analizy wariancji ANOVA (wg Witczak i in., 1994a, b)

RGQM of the upper Vistula river basin — first sampling series. Technical variance σ^2_{tech} calculated with the use of analysis of variance ANOVA (Witczak *et al.*, 1994a, b)

Lp.	Analizowana zmienna	Jednostka	N	σ^2_{tech} [%]
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	

Brak wartości σ^2_{tech} oznacza brak danych ze względu na $N < 11$ liczbę par wyników dla próbek normalnych i dublowanych

No value of σ^2_{tech} means $N < 11$ number of pairs scorses for normal and duplicate samples

Tabela 4

RMWP dorzecza górnej Wisły — pierwsza seria opróbowania. Wariancja techniczna σ^2_{tech} obliczona metodą statystyk ROBUST (wg Witczak i in., 1994a, b)

RGQM of the upper Vistula river basin — first sampling series. Technical variance σ^2_{tech} calculated with the use of ROBUST statistics (Witczak *et al.*, 1994a, b)

Lp.	Analizowana zmienna	Jednostka	N	σ^2_{tech} [%]
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	

Brak wartości σ^2_{tech} oznacza brak danych ze względu na $N < 11$ liczbę par wyników dla próbek normalnych i dublowanych

No value of σ^2_{tech} means $N < 11$ number of pairs scorses for normal and duplicate samples

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.

Po zastosowaniu elastycznego postępowania statystycznego uzyskano niższą wariancję techniczną (tab. 4). 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 fizykochemicznych. 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.

Analiza rozkładu wartości wskaźników fizykochemicznych wód podziemnych dorzecza górnej Wisły

Wyniki badań prowadzonych w sieci regionalnego monitoringu jakości wód podziemnych RMWP dorzecza górnej Wisły (Witczak i in., 1994a, b) pozwalają na uzyskanie obrazu stanu jakości wód podziemnych w zlewni górnej Wisły. W serii tej opróbowaniem objęto 167 punktów RMWP, w próbkach wody oznaczano maksymalnie 55 cech i wskaźników (55 zmiennych). Ponieważ nie w każdej z próbek oznaczano wszystkie deklarowane wskaźniki (braki danych), zmienia się liczba danych (obserwacji) w poszczególnych zbiorach.

Opis statystyczny analizowanej bazy danych (składającej się z 55 zmiennych — wskaźników fizykochemicznych wód i 167 obserwacji — opróbowanych punktów RMWP) uzyskano za pomocą programu SPSS PL for Windows v. 10.0 (SPSS, 1997a, 2000). Wykorzystano do tego celu procedurę eksploracji zbioru danych (**Analiza ▶ Opis statystyczny ▶ Eksploracja**) oraz kart kontrolnych (**Wykresy ▶ Karty kontrolne**).

Wyniki oznaczeń wskaźników chemicznych poniżej granicy oznaczalności zostały przyjęte do obliczeń jako < DL = DL (Helsel, Hirsch, 1992). W przypadku gdy wyniki < DL stanowiły ponad 20% obserwacji danej zmiennej, zmienną wyłączano z analizy — uzyskany w wyniku analizy statystycznej rozkład takiej zmiennej byłby zniekształcony przez te obserwacje (Górniak, Wachnicki, 2000).

Przeprowadzono analizę rozkładu wartości wskaźników fizykochemicznych 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; Macioszczyk, Dobrzyński, 2003; 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.

Tabela 5

Charakterystyka analizowanych zmiennych

Characteristics of analysed variables

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

N — liczba obserwacji uwzględnionych, *B* — liczba braków danych, *V* — współczynnik zmienności

N — number of observations, *B* — number of missing values, *V* — variability coefficient

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

W tabeli 5 zestawiono charakterystykę analizowanych zmiennych (wskaźników fizykochemicznych wód podziemnych w zlewni górnej Wisły) wraz ze współczynnikami zmienności obliczonymi na podstawie estymowanych parametrów rozkładu. Pełny opis statystycznej analizy rozkładu i weryfikacji każdej ze zmiennych oraz szczegółowe objaśnienia wykorzystywanych pojęć można znaleźć w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

Ostatecznie w zweryfikowanej bazie danych, na podstawie której będą prowadzone próby prognozowania jakości wód podziemnych w układzie przestrzennym pozostawiono szesnaście zmiennych. W tabeli 6 zestawiono charakterystykę tych zmiennych, podając wartość średnią $\bar{x}$ oraz parametry, na podstawie których dokonano weryfikacji.

Tabela 6

Charakterystyka zmiennych w zweryfikowanym zbiorze danych dla zlewni górnej Wisły

Variables characteristics in the verified data set

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

N — liczba obserwacji; B — liczba braków danych; $\bar{x}$ — wartość średnia; DL — laboratoryjna granica oznaczalności; PDL — praktyczna granica oznaczalności; σ^2_{tech} — wariancja techniczna, oszacowana metodą klasycznej analizy wariancji ANOVA; długa kreska (—) oznacza brak danych

N — number of values; B — number of missing values; $\bar{x}$ — mean value; DL — laboratory limit of detection; PDL — practical determination limit; σ^2_{tech} — technical variance, estimated with the use of analysis of variance ANOVA; sign — means "no data available"

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, oraz proces uczenia się, zastępujący tradycyjne programowanie.

CHARAKTERYSTYKA SIECI NEURONOWYCH

Pierwowzorem wszelkich sieci neuronowych jest mózg ludzki (szczegóły dotyczące budowy układu nerwowego człowieka, sposobu generowania połączeń można znaleźć w literaturze medycznej oraz z zakresu sieci neuronowych, np. Tadeusiewicz 1993, 1999, 2001). Rutynowo wykonuje on czynności, z którymi najszybsze komputery nie są w stanie sobie poradzić.

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.

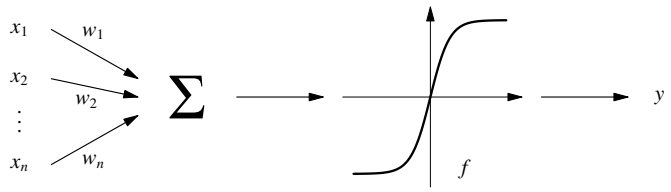


Fig. 4. Schemat sieci neuronowej (SPSS, 1997)
$x_1, \ldots, x_n$ — sygnały wejściowe; $w_1, \ldots, w_n$ — wagi połączeń; f — funkcja aktywacji neuronu; y — sygnał wyjściowy
Neural network scheme (SPSS, 1997)
$x_1, \ldots, x_n$ — input signals; $w_1, \ldots, w_n$ — weights of connections; f — activation function; y — output signal

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 (fig. 4).

Zachowanie neuronu (i całej sieci) uzależnione jest od rodzaju użytej funkcji aktywacji. Jeśli zastosowana zostanie progowa funkcja aktywacji (fig. 5a), 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 (fig. 5b).

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).

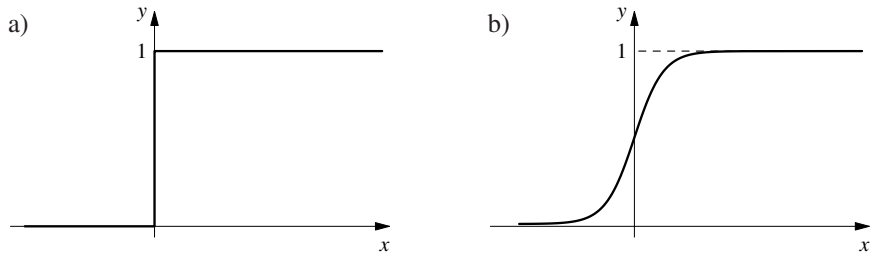


Fig. 5. Funkcje aktywacji: a) progowa; b) sigmoid (SPSS, 1997)
Activation function: a) threshold, b) sigmoid (SPSS, 1997)

Warstwa wejściowa pobiera sygnał z otoczenia (fig. 6). 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.

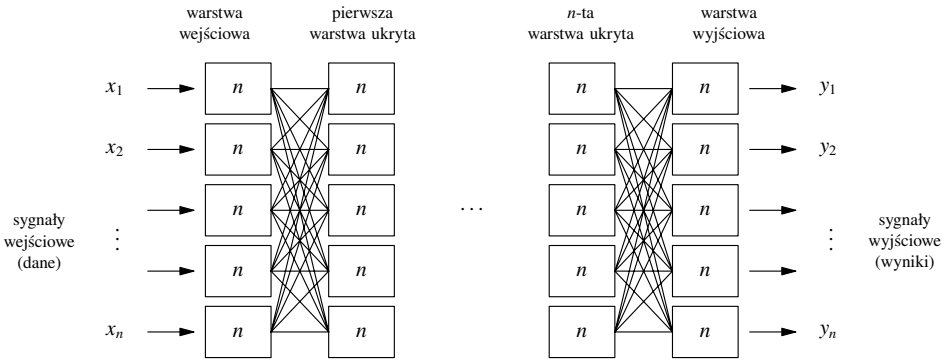


Fig. 6. Struktura sieci neuronowej (SPSS, 1997)
Neural network structure (SPSS, 1997)

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 — rodzina 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 fizykę. Spośród bardzo wielu obszarów wykorzystania sieci neuronowych, opisanych w literaturze, można wymienić m.in.: 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; nauki społeczne; kryminalistykę czy szacowanie nieruchomości (Tadeusiewicz, 1993; Broda, 2000; Falkus, Pietrzykiewicz, 2000; Hippe, 2000; Statsoft, 2000; Broda, Twardowski, 2001; Parzych 2001; Knosala i in., 2002; Kosiński, 2002).

W dziedzinie geologii jak do tej pory wykorzystywano metody rozpoznawania obrazów i sieci neuronowe, m.in. do rozpoznawania złóż (Blaschke, 1995; Kalabiński i Mastej, 1995; Kotlarczyk i in., 1995, 1997, 1999; Waksmundzki, 1995; Mastej, 2001) oraz do wyznaczania przepuszczalności skał (Jarzyna, 2000; Glazor, Broda, Twardowski, 2001). Metoda rozpoznawania obrazów (algorytm *k*-tego najbliższego sąsiada) została zastosowana przez Zamorską (1999) do prognozowania wybranych właściwości wód powierzchniowych. W 2000 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 (Lachtermacher i in., 1994; Nabagło, 1994; Świercz, 1994; Zhang i in., 1994; Zhu Mu-Lan i in., 1994; Petridis, Kehagias, 1998; Pociask-Karteczka, 1999; Lula, 2001). 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 badania dotyczące wykorzystania sieci neuronowych do prognozowania zmian jakości wód w układzie czasowym (Kmiecik 2000; Szczepańska, Kmiecik, 2000, 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*.

ZASTOSOWANIE SIECI NEURONOWYCH DO PREDYKCJI I KLASYFIKACJI

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

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

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 (fig. 7).

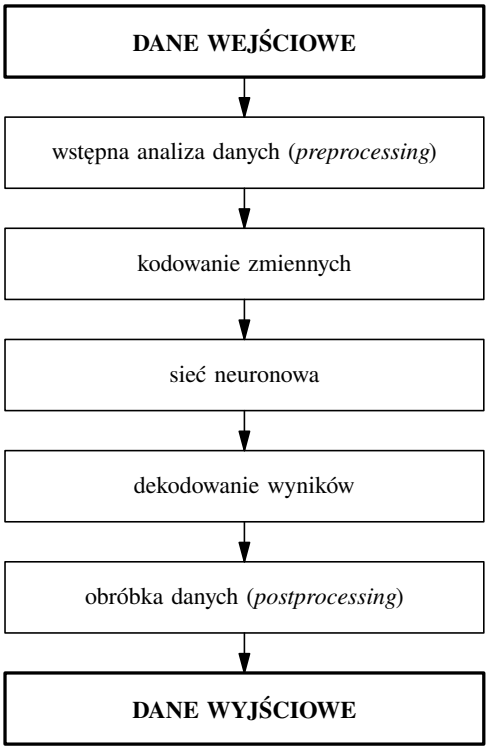


Fig. 7. Schemat analizy danych za pomocą sieci neuronowych (SPSS, 1997)
Scheme of data analysis with the use of neural networks (SPSS, 1997)

Pierwszym krokiem w modelowaniu danych za pomocą sieci neuronowych jest poddanie ich wstępnej analizie (*preprocessing*), stosując wiele technik analitycznych w celu wyselekcjonowania 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 podlega procesowi dalszej obróbki, tzw. *postprocessingu*.

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 wyrazić w postaci liczby dni, jaka minęła od pewnej daty odniesienia (*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 (*training*) — inaczej uczący, prezentowany sieci w trakcie uczenia i służący do modyfikacji parametrów (zbiór, na którym sieć uczy się odwzorowania danych);
- walidacyjny (*validation*) — inaczej weryfikacyjny, zbiór pozwalający na monitorowanie procesu uczenia sieci;
- testowy (*test*) — służący do stwierdzenia poprawności zbudowanego modelu.

Rodzaj zadania, jakie sieć ma modelować ma wpływ na wybór techniki podziału danych na podzbiory: treningowy, walidacyjny i testowy. Jeśli 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).

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.

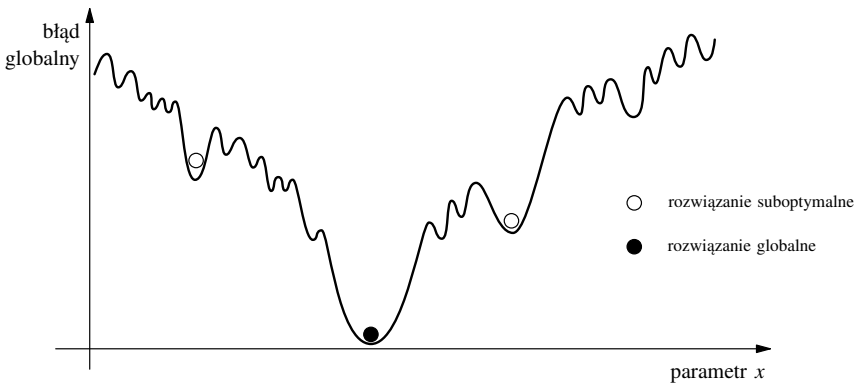


Fig. 8. Ilustracja obszaru błędu (SPSS, 1997)
Illustration of error domain (SPSS, 1997)

Celem analizy prowadzonej z wykorzystaniem sieci neuronowych jest znalezienie rozwiązania globalnego (fig. 8) 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 był równy zeru.

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 zagadnień:

- 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 oparta na informacjach historycznych tej samej cechy);
- segmentacji danych (podział dużych baz danych na klastry na podstawie podobieństwa 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).

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

Perceptron wielowarstwowy (*Multi-Layer Perceptron*, MLP). Model perceptronu wielowarstwowego 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. Jest to sieć neuronowa oparta na oryginalnym modelu prostego perceptronu, z dodatkowymi warstwami neuronów ukrytych pomiędzy warstwą wejściową i wyjściową, pozwala zatem na dokładne odzwierciedlenie nieliniowości danych.

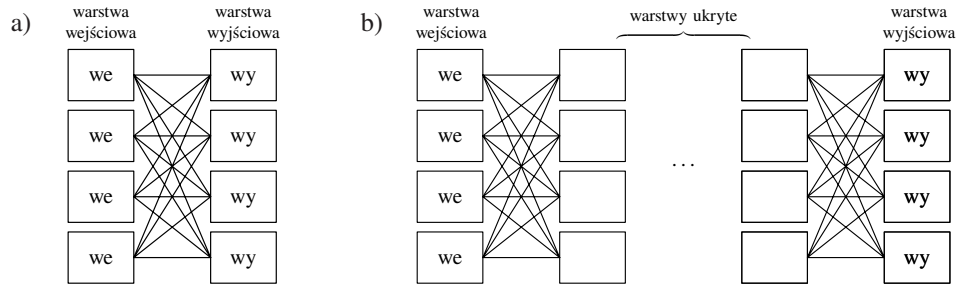


Fig. 9. Model zwykłego (pojedynczego) perceptronu (a) i perceptronu wielowarstwowego (b) (SPSS, 1997)
Model of a) simple perceptron and b) multilayer perceptron (SPSS, 1997)

Zwykły perceptron składa się z warstwy wejściowej i wyjściowej, bez warstw ukrytych. Każdy neuron w warstwie wejściowej jest połączony z każdym neuronem w warstwie wyjściowej. W trakcie uczenia sieci dopasowywane są „odległości" między neuronami (fig. 9a). Wynik dla neuronu w perceptronie jest iloczynem wartości wprowadzonych na wejściu i odpowiednich wag. Pobierając 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 wykorzystaniem funkcji aktywacji do modyfikacji wejścia neuronu (fig. 9b). 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 następująco: najpierw są inicjowane 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, gdy otrzymany zostanie wynik w warstwie wyjściowej (SPSS, 1997). Następnie mierzona jest różnica między 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, gdy 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 (fig. 8), 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 między 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 rozbudowywania sieci ponad jedną warstwę ukrytą (Tadeusiewicz, 1993).

Na figurze 10 przedstawiono schematycznie problem klasyfikacji za pomocą sieci MLP dla dwóch cech wejściowych i dwóch klas danych wyjściowych.

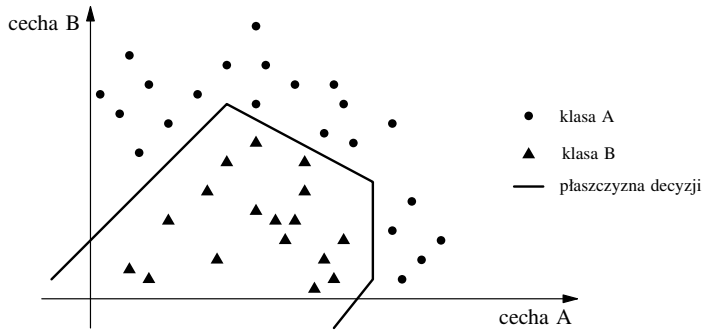


Fig. 10. Płaszczyzna decyzji w sieci typu MLP — problem klasyfikacji z dwiema cechami wejściowymi i dwiema klasami danych wyjściowych (SPSS, 1997)

Decision surface in the MLP network — classification problem with two input features and two output classes (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 i 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.

Radialna funkcja bazowa (*Radial Basis Function*, RBF). Inną siecią z grupy sieci nadzorowanych jest sieć z radialną funkcją bazową RBF. Ta struktura nie konstruuje płaszczyzny decyzji w przestrzeni danych wejściowych, tak jak sieć MLP — dane są klastrowane przez kilka funkcji bazowych (fig. 11).

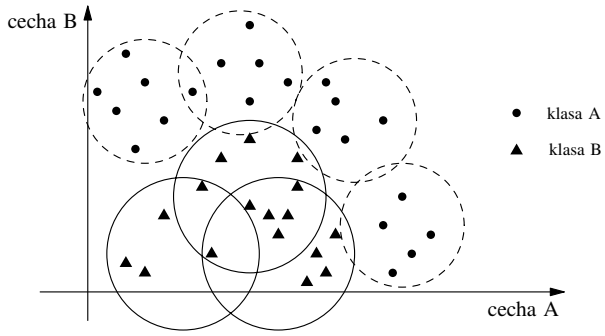


Fig. 11. Płaszczyzna decyzji w sieci typu RBF — problem klasyfikacji z dwiema cechami wejściowymi i dwiema klasami danych wyjściowych (SPSS, 1997)

Decision surface in the RBF network — classification problem with two input features and two output classes (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 sieci MLP a ponadto 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: liczbę centrów wymaganą do dokładnego modelowania danych, pozycjonowanie centrów i 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 (*overfitting*) i błędnych uogólnień. Pozycjonowanie centrów zależy od sposobu, w jaki inicjowane są 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).

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 przez 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).

Walidacja modelu sieci neuronowej

Celem procesu uczenia sieci neuronowej jest minimalizacja błędu między 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, niewykorzystanych 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 (fig. 12).

W początkowych fazach procesu uczenia błędy: uczenia i walidacyjny, pozostają w stałym stosunku, w chwili 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 (fig. 8) można stosować różne warunki początkowe (fig. 13).

Celem algorytmu uczącego jest doprowadzenie systemu do stanu, w którym osiągnie on najmniejszy błąd. W przypadku przedstawionym na figurze 13a jest duża szansa, że osiągnięte rozwiązanie będzie rozwiązaniem suboptymalnym.

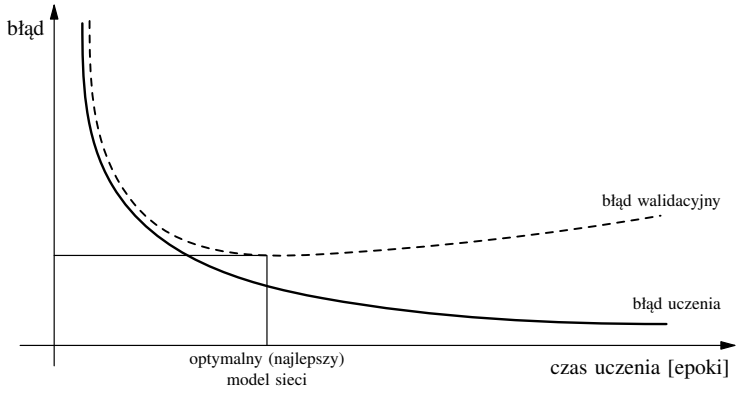


Fig. 12. Walidacja modelu sieci neuronowej (SPSS, 1997)
Validation of neural network model (SPSS, 1997)

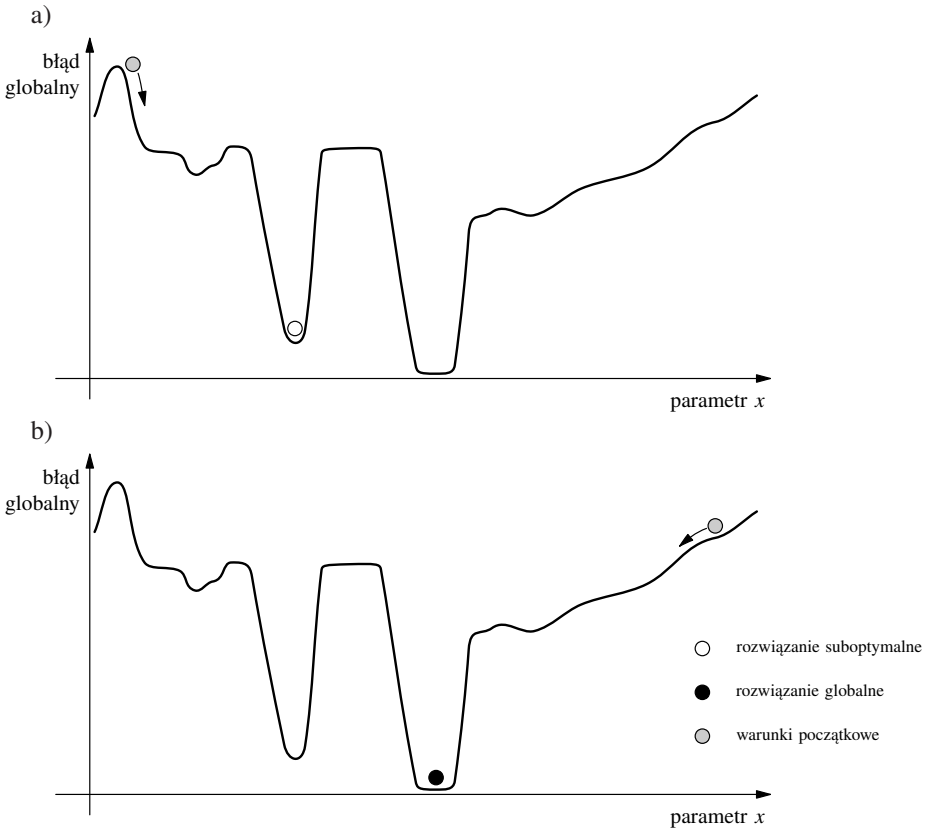


Fig. 13. Ilustracja efektu zmiany warunków początkowych (SPSS, 1997)
Illustration of the effect of changing the initial conditions (SPSS, 1997)

W sytuacji przedstawionej na figurze 13b zmienione zostały początkowe wartości wag i wynik końcowy będzie prawdopodobnie rozwiązaniem optymalnym. Poprzez zmianę warunków początkowych przeprowadzona została eksploracja powierzchni błędu.

Doświadczony eksperymentator sam buduje sztuczną sieć neuronową, określając jej architekturę poprzez podanie rozmiaru warstwy wejściowej, liczby i rozmiaru warstw ukrytych oraz liczby neuronów wyjściowych. Dodatkowo może on zdefiniować sprzężenie w sieci, wagi oraz funkcje aktywacji neuronów. Mniej doświadczony badacz może korzystać z dostępnej w programach komputerowych opcji automatycznego generowania sieci optymalnej. Nie ma tu więc znaczenia liczba architektur budowanych sieci, liczy się natomiast krotność powtórzeń przetwarzania zbioru za pomocą określonej sieci:

- w przypadku zbiorów o liczebności mniejszej niż 100 — algorytm *leaving-one-out*;
- dla zbiorów o liczebności większej niż 100 — algorytm *ten-fold-cross validation*;
- dla zbiorów o liczebności ponad 5000 — walidacja przez podział.

W każdym przypadku należy po przeprowadzeniu odpowiedniej liczby prób uśrednić wyniki. Szczegóły dotyczące algorytmów walidacji modelu sieci neuronowej można znaleźć w literaturze (np. Tadeusiewicz, 1993). W dokumentacji do programu Neural Connection czytamy, że 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).

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

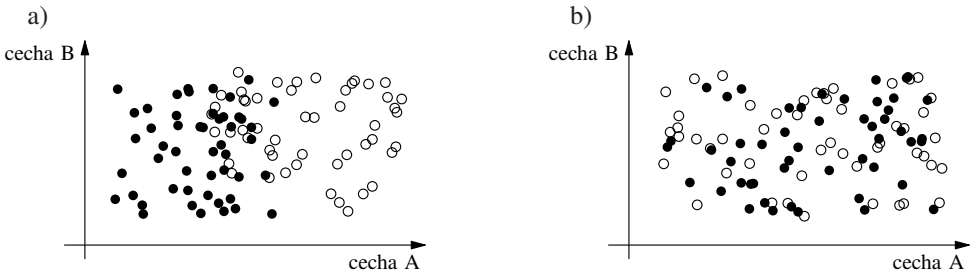


Fig. 14. Sposoby wyboru zbiorów treningowego i testowego: a) błędny; b) poprawny (SPSS, 1997)

Illustration of recutting the training and test files: a) poor representation, b) better representation (SPSS, 1997)

Jeśli dane były zbierane w różnym czasie, obserwacje do zbioru testowego powinny być tak dobrane, by reprezentowały 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 (fig. 14).

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

PROGRAMY KOMPUTEROWE DO BUDOWANIA MODELI SIECI NEURONOWYCH

Większość programów komputerowych do tworzenia modeli sieci neuronowych pozwala użytkownikowi w bardzo łatwy sposób przeprowadzić analizę danych. 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 (**http://www.it.uom.gr/pdp/DigitalLib/Neural/Neu_soft.htm**) można znaleźć wykaz (i krótki opis) dostępnego na wszystkie platformy systemowe oprogramowania (komercyjnego i *freeware*) dotyczącego sieci neuronowych. Na liście tej znajdują się m.in. takie programy komercyjne, jak: Adaptive Logic Network, DataEngine, ECANSE — Environment for Computer Aided Neural Software Engineering, MATLAB: Neural Network Toolbox, Neural Bench, 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, Hyperplane Animator, Mume, Neocognitron, Nenet, Net II, NETS — Network Execution and Training Simulator, Neural Shell, Pittnet, Pygmalion, WinNN, Qnet2000 czy Neurooffice.

Próby tworzenia modeli sieci neuronowych dla potrzeb rozprawy doktorskiej (Kmiecik, 2001) i 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, zarządzanie danymi, modelowanie, raportowanie, tworzenie diagramów przepływu danych, działa na platformach Win98, WinNT i 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; możliwość bezpośredniej wymiany danych) wybrano program Neural Connection v. 2.1 (SPSS, 1997, 1999, 2000).

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

Program Neural Connection v. 2.1

Program Neural Connection pozwala na budowanie modeli sieci neuronowych do różnego rodzaju zastosowań i ma niewielkie wymagania sprzętowe i systemowe (SPSS, 1997):

- komputer PC z procesorem co najmniej 386;
- system operacyjny Microsoft Windows 95, 98, NT 4.0 lub XP;
- 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.

Przykłady rozwiązywania zagadnień predykcji i klasyfikacji w układzie przestrzennym za pomocą programu Neural Connection znajdują się w kolejnym rozdziale. Szczegóły dotyczące programu (opis opcji, procedur, przykłady zastosowań, itp.) można znaleźć w dokumentacji (SPSS, 1997, 1999), w systemie pomocy do programu (*Help*) oraz w pracy (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

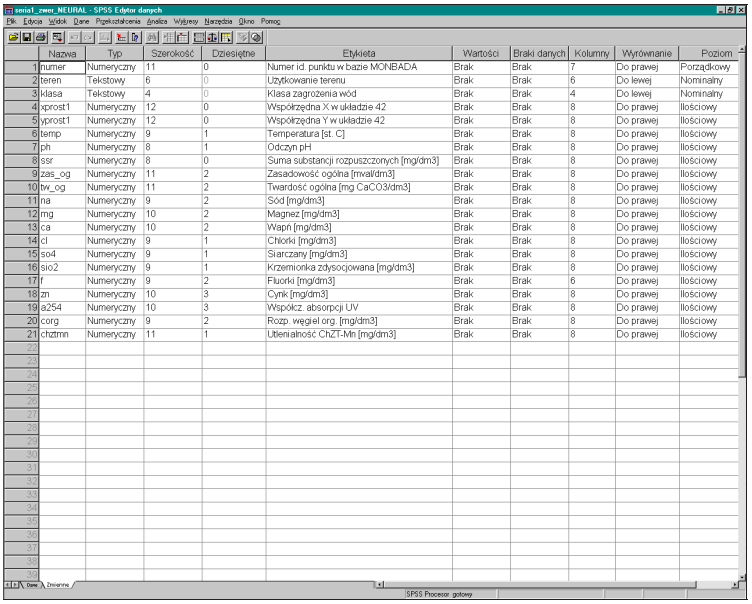
PROGNOZOWANIE ZMIAN JAKOŚCI WÓD PODZIEMNYCH W UKŁADZIE PRZESTRZENNYM

Do prognozowania zmian jakości wód podziemnych dorzecza górnej Wisły w układzie przestrzennym wykorzystano dane uzyskane w pierwszej serii opróbowania sieci RMWP (okres mokry, V–IX 1993).

W zweryfikowanej bazie danych (omówionej wcześniej) pozostało 16 zmiennych — wskaźników fizykochemicznych 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 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 zapisano do pliku w formacie SPSS, pod nazwą *seria1_zwer_NEURAL.sav*, plik ten można znaleźć na stronie http://galaxy.agh.edu.pl/~ek. W pliku jest zatem **21 zmiennych** (16 wskaźników fizykochemicznych wód i 5 zmiennych typu opisowego — fig. 15) opisujących **167 punktów RMWP** (167 wierszy danych — fig. 16).

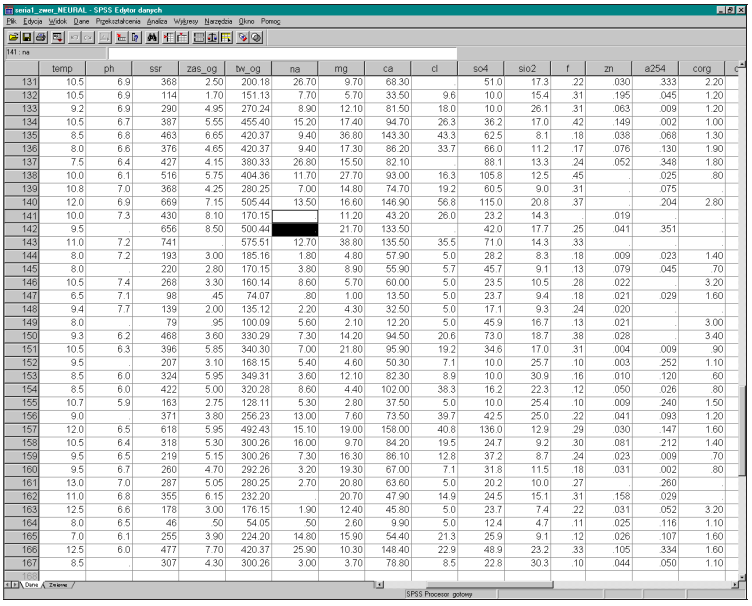
Na tak przygotowanej bazie danych zostały przeprowadzone próby **predykcji** wskaźników fizykochemicznych wód na podstawie współrzędnych punktu monitoringowego oraz **klasyfikacji** punktu monitoringowego (na podstawie wartości wyników oznaczeń wskaźników fizykochemicznych) do obszaru o określonym użytkowaniu terenu. Próby te były prowadzone 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	twog	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

Fig. 15. Zweryfikowana baza danych — podgląd zmiennych

Verified database — variables



	temp	ph	ssr	zas_og	twog	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

Fig. 16. Zweryfikowana baza danych — podgląd danych. Zaznaczone są przykłady braków danych w zmiennej *sód* [mg/dm^3]

Verified database — variables. Selected cells are missing values in the variable sodium [mg/dm^3]

- **Wariant 1.** Plik z danymi reprezentującymi wszystkie klasy zagrożenia wód (AB, C, D) i sposoby 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) medianą ze wszystkich obserwacji (medianę wybrano ze względu na asymetryczne rozkłady badanych zmiennych). W tym celu wykorzystano opcję programu SPSS **Przekształcenia ▶ Zastąp braki danych**. W pliku dla wariantu 1 jest zatem 21 zmiennych (16 wskaźników fizykochemicznych i 5 zmiennych opisowych) i 167 obserwacji (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 wyłączono z analizy punkty RMWP o klasach zagrożenia C i D. Plik ten ma taką samą konfigurację jak plik w wariancie 1, składa się z 21 zmiennych (16 wskaźników fizykochemicznych i 5 zmiennych opisowych), 151 obserwacji (151 punktów RMWP).
- **Wariant 3.** Plik z danymi reprezentującymi klasę zagrożenia wód AB z ograniczoną liczbą zmiennych (wskaźników fizykochemicznych).
  W celu sprawdzenia, czy na jakość uzyskiwanych prognoz ma wpływ operacja zastępowania znacznej liczby braków danych medianą, w wariancie trzecim testowano plik zawierający punkty o klasie zagrożenia AB, ale ograniczony do 11 zmiennych (6 wskaźników fizykochemicznych i 5 zmiennych opisowych). Spośród zweryfikowanych wskaźników fizykochemicznych do pliku wybrano te, w których wystąpiła najmniejsza ($\leqslant 5$) liczba braków danych (tab. 6): 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] i krzemionka zdysocjowana [mg/dm^3]. Następnie z pliku wyłączono (usunięto) wszystkie obserwacje/wiersze z brakami danych — w zbiorze pozostało więc 143 obserwacje (143 punkty RMWP).

W tabeli 7 zestawiono konfiguracje plików dla poszczególnych wariantów.

Tabela 7

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

Variants of files used in predictions of spatial groundwater quality changes in the upper Vistula river basin

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

PROGNOZOWANIE WARTOŚCI WSKAŹNIKÓW JAKOŚCI WÓD
NA PODSTAWIE WSPÓŁRZĘDNYCH PUNKTU MONITORINGOWEGO

W celu stwierdzenia czy na podstawie wartości 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 — wyniki oznaczeń wskaźników fizykochemicznych wód.

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

Przygotowany zgodnie z wariantem pierwszym (tab. 7) plik danych *zbior01.sav* (ten plik, wszystkie pliki z danymi, pliki wynikowe oraz pliki z modelami omawianych sieci neuronowych można znaleźć na stronie http://galaxy.agh.edu.pl/~ek) wczytano wprost do programu Neural Connection, uruchamiając z programu SPSS opcję **Analiza ▶ Neural Connection**, a następnie dokonano konfiguracji zmiennych (fig. 17).

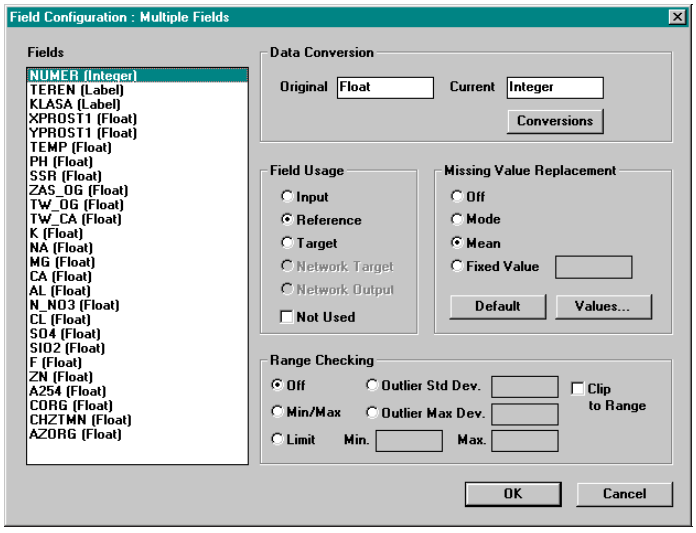
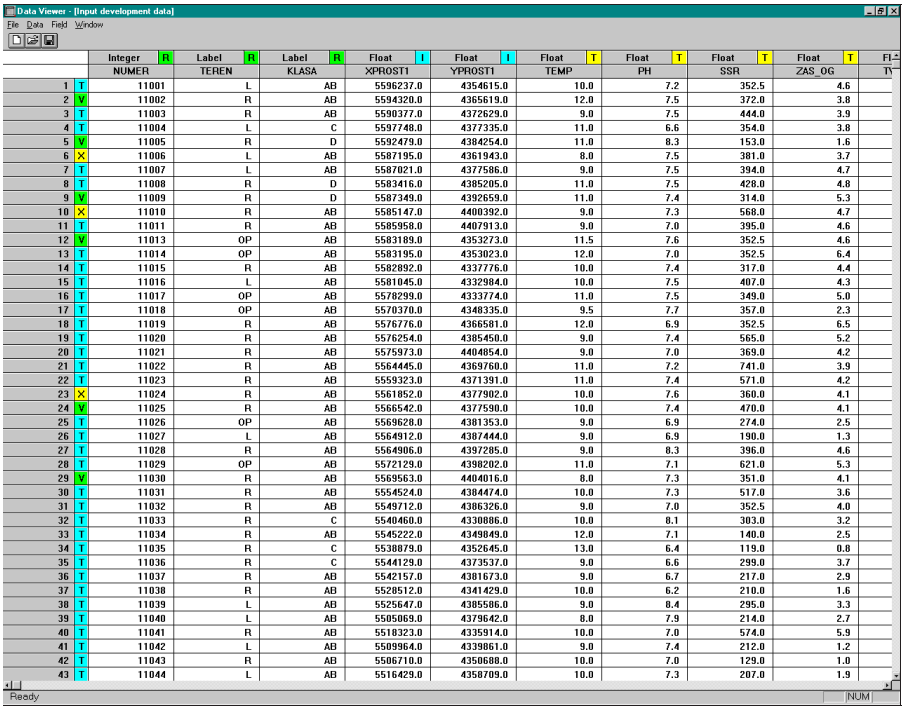


Fig. 17. Ekran konfiguracji zmiennych
Screen of variables configuration

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) — wartości wskaźników fizykochemicznych wód (fig. 18).



Data Viewer - [Input development data]

File Data Field Window

		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	

Ready NUM

Fig. 18. Ekran podglądu danych wejściowych

R — zmienne typu opisowego; I — zmienne typu wejściowego; T — zmienne docelowe

Input data

R — reference variables; I — input variables; T — target variables

Następnie przystąpiono do budowy modelu sieci. W celu wyboru optymalnej (dającej najlepsze rezultaty prognoz) struktury sieci testowano różne modele z grupy sieci nadzorowanych, dostępne w programie Neural Connection: wielowarstwowy perceptron MLP, radialną funkcję bazową RBF i sieć Bayesa.

Sieć typu MLP. Schemat zbudowanej sieci neuronowej przedstawia figura 19.

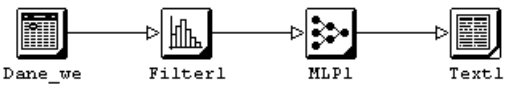


Fig. 19. Schemat sieci MLP do prognozowania wartości wskaźników jakości wód podziemnych na podstawie współrzędnych punktu monitoringowego

Scheme of MLP network for prediction of groundwater quality indicators values on the base of monitoring sites coordinates

Narzędzie do filtracji (fig. 20) pozwala na ograniczanie liczebności zbioru wejściowego poprzez „odfiltrowanie danych" (np. „przycięcie" 5% obserwacji z góry i dołu obserwowanego zakresu zmiennej) lub wyłączenie danej zmiennej z analizy. Umożliwia ono także dokonywanie przekształceń danych (np. operacje logarytmowania) i analizę rozkładu zmiennej.

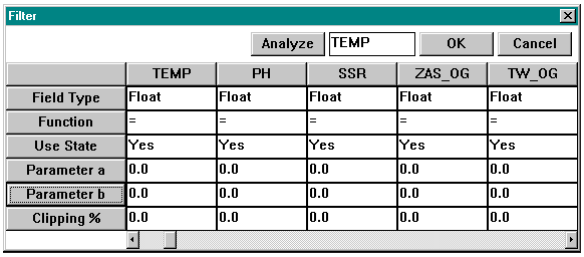


Fig. 20. Narzędzie do filtracji danych

Data filtration tool

Zbiór danych wejściowych został podzielony na podzbiory (fig. 21) 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).

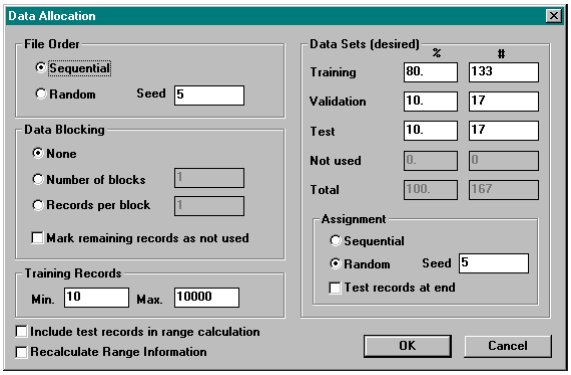


Fig. 21. Ekran podziału danych wejściowych na podzbiory

Input data allocation screen

Pierwszą próbę „uczenia" sieci przeprowadzono przy domyślnych ustawieniach opcji konfiguracji modułu MLP (fig. 22):

- 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*; możliwe do wyboru są jeszcze funkcje sigmoid i liniowa) i jednostajny rozkład wag neuronów (*Uniform*; opcjonalnie można wybrać rozkład normalny);
- opcja *Use Best Network* umożliwia automatyczny wybór optymalnej sieci (o najlepszych parametrach).

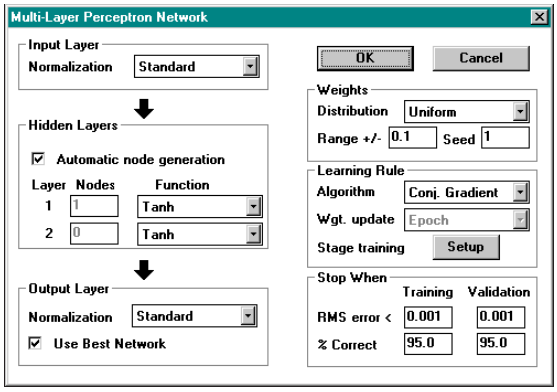


Fig. 22. Okno konfiguracji struktury MLP

MLP structure configuration window

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).

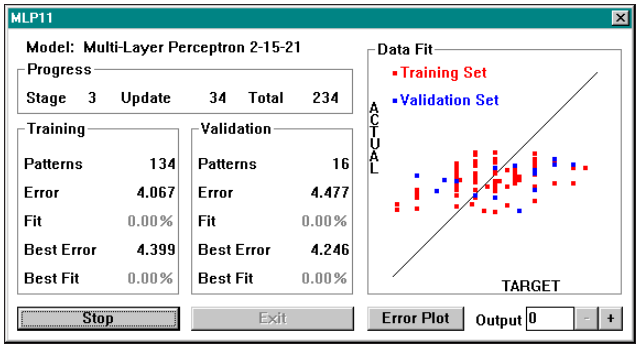


Fig. 23. Okno „uczenia się" struktury MLP

MLP structure learning window

Proces „uczenia się" sieci (fig. 23) 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:

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

gdzie:

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

Wyniki „uczenia sieci" dla zbioru treningowego zostały zapisane do zbioru tekstowego *mlp01.nno* (fig. 24).

```
! 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 %
       ...
```

Fig. 24. Fragment zbioru wynikowego *mlp01.nno*

Excerpt from output data file mlp01.nno

W zbiorze tym znajdują się oryginalne zmienne prognozowane, zmienne uzyskane jako odpowiedź sieci i parametry określające jakość prognoz dla każdej ze zmiennych — błąd średniokwadratowy (*RMS Error*), odchylenie przeciętne MA (*Mean Absolute*):

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

i średni błąd względny w procentach, liczony jako moduł różnicy między wartością prawdziwą a wartością prognozy MA% (*Mean Absolute %*) (oznaczenia jak we wzorze 1):

$$\mathrm{MA\%} = \mathrm{MA} \cdot \frac{N}{\sum_{i=1}^{N} z_i} \cdot 100\% \tag{3}$$

Ten ostatni parametr podawany jest w zestawieniach, w celu porównania prognoz uzyskanych za pomocą sieci o różnej konfiguracji.

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 (*mlp01.nno–mlp11.nno*) i uzyskane dla zbioru treningowego wyniki prognoz wartości wskaźników chemicznych wód znajdują się w pracy (Kmiecik, 2001) na stronie internetowej http://galaxy.agh.edu.pl/˜ek.

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 fizykochemiczne). Nie stwierdzono związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników.

Najlepsze wyniki uzyskano dla struktury 2–12–16 (liczba neuronów w warstwie wejściowej–liczba neuronów w warstwie ukrytej–liczba neuronów w warstwie wyjściowej; 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.

Sieć typu RBF. Schemat zbudowanej sieci neuronowej typu RBF przedstawia figura 25.

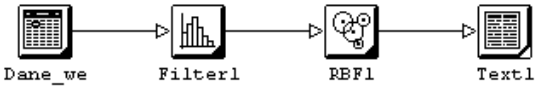


Fig. 25. Schemat sieci RBF do prognozowania wskaźników jakości wód podziemnych na podstawie współrzędnych punktu monitoringowego

Scheme of RBF network for prediction of groundwater quality indicators values on the base of monitoring sites coordinates

Podział zbioru danych wejściowych był taki sam 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 (fig. 26):

- standardowa normalizacja warstw wejściowej i wyjściowej;
- liczba centrów: 5;
- miara odległości błędu: Euklidesowa (*Euclidean*);
- funkcja nieliniowa: *Spline* (funkcja ta wyraża się wzorem $d^2 \log d$, gdzie: d — odległość od centrum; β — parametr funkcji);
- rozkład centrów: w oparciu o dane (*Sample*) — centra funkcji bazowych znajdują się w wybranych punktach danych;
- optymalizacja: zwiększanie liczby centrów o 5 aż do 50;
- warunki zakończenia procesu uczenia: zmiana błędu −0,01% w ciągu 5 cykli (epok).

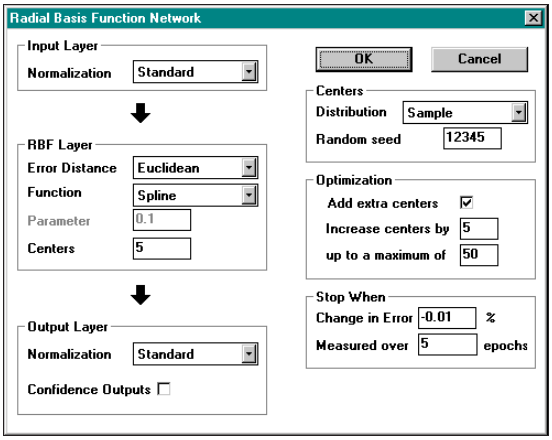


Fig. 26. Okno konfiguracji struktury RBF

Window for RBF structure configuration

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 (*rbf01.nno–rbf15.nno*) i uzyskane wyniki prognoz zestawione są w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

Błędy względne prognoz wartości wskaźników fizykochemicznych 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.

Najlepsze wyniki prognoz (najmniejsze błędy względne prognoz) uzyskano dla modelu, którego wyniki zapisane są w pliku *rbf15.nno* (odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0,2; liczba centrów: 5; rozkład centrów: próbny (*Trial*)).

Sieć typu Bayesa. Schemat zbudowanej sieci neuronowej typu Bayesa przedstawia figura 27.

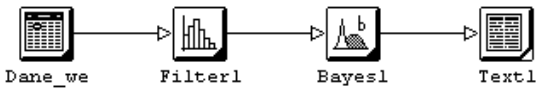


Fig. 27. Schemat sieci Bayesa do prognozowania wskaźników jakości wód podziemnych na podstawie współrzędnych punktu monitoringowego

Scheme of Bayesian network for prediction of groundwater quality indicators values on the base of monitoring sites coordinates

Podział zbioru danych wejściowych zachowano jak w przypadku sieci MLP i RBF: podzbiór treningowy (80% wszystkich obserwacji — 133 obserwacje), walidacyjny (10% — 17 obserwacji) i testowy (17 obserwacji). Wybrano opcję wykorzystania przez sieć zbioru walidacyjnego w procesie jej „uczenia się". 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.

Wybór najlepszego modelu sieci. Porównując wyniki prognoz uzyskanych za pomocą modeli MLP i RBF, najlepszym modelem (najmniejsze błędy względne prognoz wartości badanych wskaźników) okazał się model sieci RBF, którego wyniki zapisane są w pliku *rbf15.nno*.

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 (fig. 28).

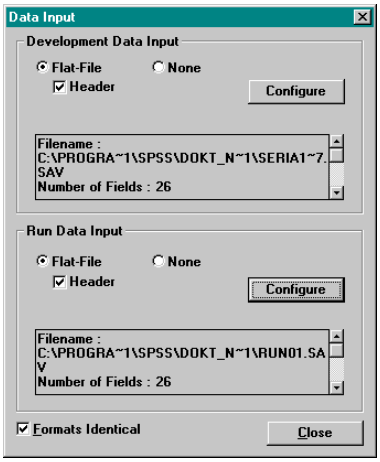


Fig. 28. Okno do konfiguracji pliku roboczego (*run*)

Run file configuration window

Wyniki prognoz wartości wskaźników fizykochemicznych 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* (fig. 29, 30).

W programie SPSS obliczono błędy względne prognoz *B* dla każdego z prognozowanych wskaźników:

$$B = \frac{x_{obs} - x_{progn}}{x_{obs}} \cdot 100\% \tag{4}$$

gdzie:

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

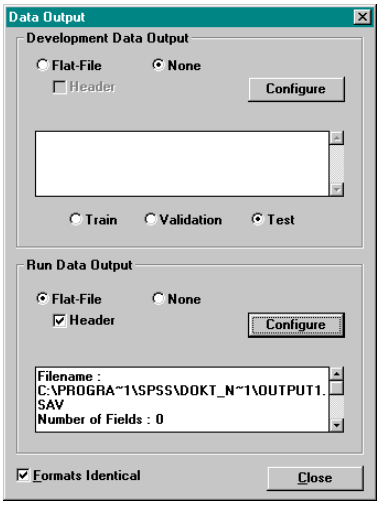
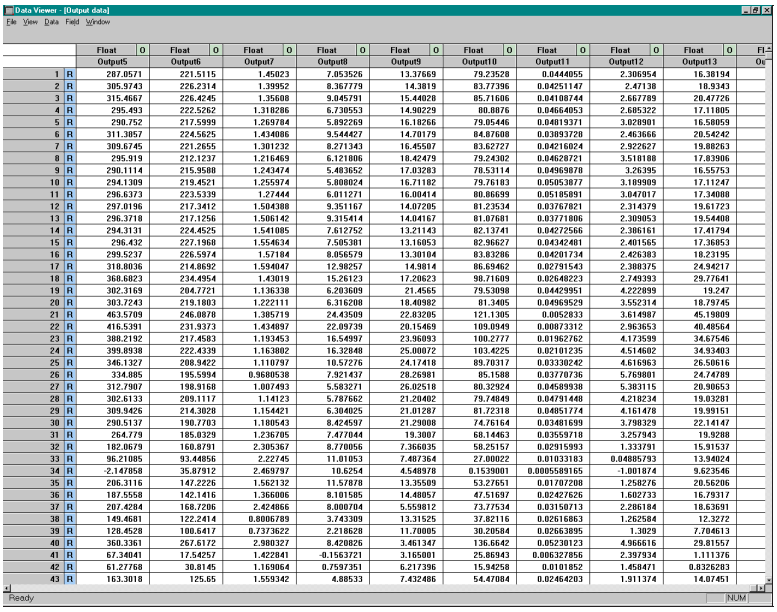


Fig. 29. Okno do konfiguracji zapisu wyników prognoz dla pliku roboczego do pliku zewnętrznego w formacie SPSS

Configuration of run data output window to the file in SPSS format



	Output5	Output6	Output7	Output8	Output9	Output10	Output11	Output12	Output13
1	287.0571	221.5115	1.45023	7.053526	13.37669	79.23528	0.0444055	2.306954	16.38194
2	305.9743	226.2314	1.39952	8.367779	14.3819	83.77396	0.04251147	2.47138	18.9343
3	315.4667	226.4245	1.35608	9.045791	15.44028	85.71606	0.04108744	2.667789	20.47726
4	295.493	222.5262	1.318286	6.730553	14.90229	80.8876	0.04664053	2.685322	17.11805
5	290.752	217.5999	1.269784	5.892269	16.10266	79.05446	0.04019371	3.020901	16.58059
6	311.3857	224.5625	1.434086	9.544427	14.70179	84.87608	0.03893728	2.463666	20.54242
7	309.6745	221.2655	1.381232	8.271343	16.45507	83.62727	0.04216024	2.922627	19.88263
8	295.919	212.1237	1.216469	6.121006	18.42479	79.24302	0.04628721	3.518188	17.83906
9	290.1114	215.9588	1.243474	5.403652	17.03283	78.53114	0.04969878	3.26395	16.55753
10	294.1309	219.4521	1.255574	5.808024	16.71182	79.76183	0.05053877	3.189909	17.11247
11	296.6373	223.5339	1.27444	6.011271	16.00414	80.86699	0.05185891	3.047017	17.34008
12	297.0196	217.3412	1.504388	9.351167	14.07205	81.23534	0.03767821	2.314379	19.61723
13	296.3718	217.1256	1.506142	9.315414	14.04167	81.07681	0.03771806	2.309053	19.54408
14	294.3131	224.4525	1.541085	7.612752	13.21143	82.13741	0.04272566	2.386161	17.41794
15	296.432	227.1968	1.554634	7.505381	13.16053	82.96627	0.04342481	2.401565	17.36853
16	299.5237	226.5974	1.57184	8.056579	13.30104	83.83286	0.04201734	2.426383	18.23195
17	310.8036	214.8692	1.594047	12.98257	14.9814	86.69462	0.02791543	2.388375	24.94217
18	360.6823	234.4954	1.43019	15.26123	17.20623	98.71609	0.02648223	2.749393	29.77641
19	302.3169	204.7721	1.136338	6.203609	21.4565	79.53098	0.04429951	4.222899	19.247
20	303.7243	219.1803	1.222111	6.316208	18.40982	81.3405	0.04969529	3.552314	18.79745
21	463.5709	246.0878	1.385719	24.43509	22.83205	121.1305	0.0052833	3.614987	45.19809
22	416.5391	231.9373	1.434897	22.09739	20.15469	109.0949	0.00873312	2.963653	40.48564
23	388.2192	217.4583	1.193453	16.54997	23.96093	100.2777	0.01962762	4.173599	34.67546
24	399.8938	222.4339	1.163802	16.32848	25.00072	103.4225	0.02101235	4.514602	34.93403
25	346.1327	208.9422	1.110797	10.57276	24.17418	89.70317	0.03330242	4.616963	26.50616
26	334.885	195.5994	0.9680538	7.921437	28.26981	85.1588	0.03770736	5.769801	24.74789
27	312.7907	198.9168	1.007493	5.583271	26.02518	80.32924	0.04589938	5.383115	20.90653
28	302.6133	209.1117	1.14123	5.787662	21.20402	79.74849	0.04791448	4.218234	19.03281
29	309.9426	214.3028	1.154421	6.304025	21.01287	81.72318	0.04851774	4.161478	19.99151
30	290.5137	198.7703	1.180543	8.424597	21.29008	74.76164	0.03481699	3.798329	22.14147
31	264.779	185.0329	1.236705	7.477044	19.3007	68.14463	0.03559718	3.257943	19.9208
32	182.0679	160.8791	2.305367	8.770056	7.366035	58.25157	0.02915993	1.333791	15.91537
33	96.21085	93.44856	2.22745	11.01053	7.487364	27.00022	0.01033183	0.04885793	13.94024
34	-2.147858	35.87912	2.469797	10.6254	4.548978	0.1539001	0.0005589165	-1.001874	9.623546
35	206.3116	147.2226	1.562132	11.57878	13.35509	53.27651	0.01707208	1.258276	20.56206
36	187.5558	142.1416	1.366006	8.101585	14.48057	47.51697	0.02427626	1.602733	16.79317
37	207.4284	168.7206	2.424866	8.000704	5.559812	73.77534	0.03150713	2.286184	18.63691
38	149.4681	122.2414	0.8006789	3.743309	13.31525	37.82116	0.02616863	1.262584	12.3272
39	128.4528	108.6417	0.7373622	2.218628	11.70005	30.20584	0.02663895	1.3029	7.704613
40	360.3361	267.6172	2.980327	8.420826	3.461347	136.6642	0.05230123	4.966616	29.81557
41	67.34041	17.54257	1.422841	-0.1563721	3.165001	25.86943	0.006327856	2.397934	1.111376
42	61.27768	30.8145	1.169064	0.7597351	6.217396	15.94258	0.0101852	1.458471	0.8326283
43	163.3018	125.65	1.559342	4.88533	7.432486	54.47084	0.02464203	1.911374	14.07451

Fig. 30. Wyniki prognoz wartości wskaźników fizykochemicznych na podstawie współrzędnych punktu monitoringowego uzyskane z modelu RBF dla pliku roboczego (*run*)

Results of physicochemical indicators predictions on the base of monitoring sites coordinates obtained from RBF model for run file

Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych. Pełny raport z tej części analizy znajduje się w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek), w pliku *output1.spo*.

W tabeli 8 zestawione są średnie błędy prognoz *B* (obliczone wg wzoru 4) dla analizowanych zmiennych — wartości wskaźników fizykochemicznych oraz współczynniki korelacji wartości obserwowanych i prognozowanych.

Tabela 8

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

Analysis of predictions of groundwater quality indicators values on the base of monitoring sites coordinates — RBF network, run file, points of AB, C, D class (16 target variables)

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

Średni błąd względny prognoz wartości 16 wskaźników fizykochemicznych wód kształtuje się na niskim poziomie, rzędu od setnych procenta do kilkunastu procent. Najmniejszy błąd prognoz cechuje sumę substancji rozpuszczonych (−0,12%; fig. 31a) i odczyn pH (0,31%). W przypadku fluorków średni błąd względny prognoz osiąga wartość −50% (fig. 32a), a przypadku krzemionki wartość maksymalną 108,83%.

Rozkłady błędów względnych prognoz charakteryzuje w przypadku niektórych wskaźników chemicznych bardzo duży rozrzut — np. dla wapnia od −6500 do 7500%, czy w przypadku krzemionki od −1000 do 19 000%. 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.

Współczynniki korelacji wartości obserwowanych z prognozowanymi mieszczą się w zakresie 0,383–0,762. Nie są to więc prognozy dobrej jakości.

W przypadku cynku, który charakteryzował się najniższą precyzją ($\sigma^2_{tech} = 86,85\%$) 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.

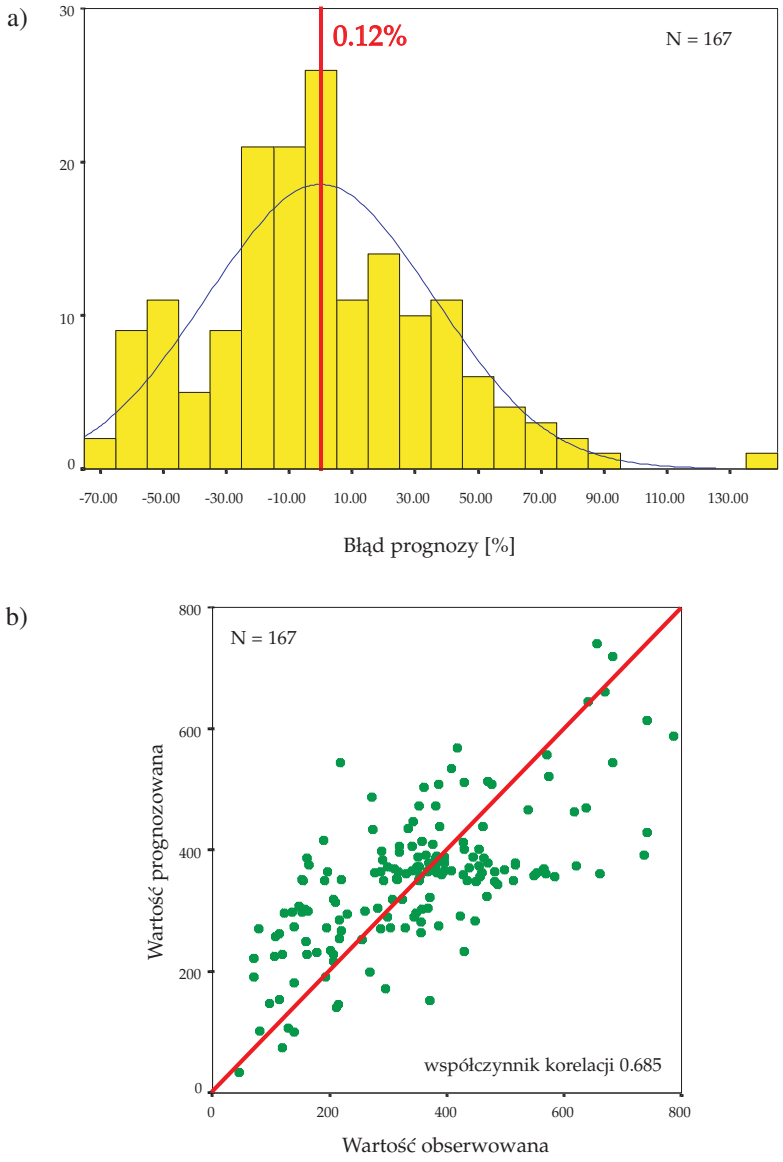


Fig. 31. Prognozowanie stężeń sumy substancji rozpuszczonych [mg/dm^3] na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB, C, D

Predictions of total dissolved solids concentrations [mg/dm^3] on the base of monitoring sites coordinates: a) frequency histogram of relative prediction error, b) scatterplot of observed and predicted values; points of AB, C, D class

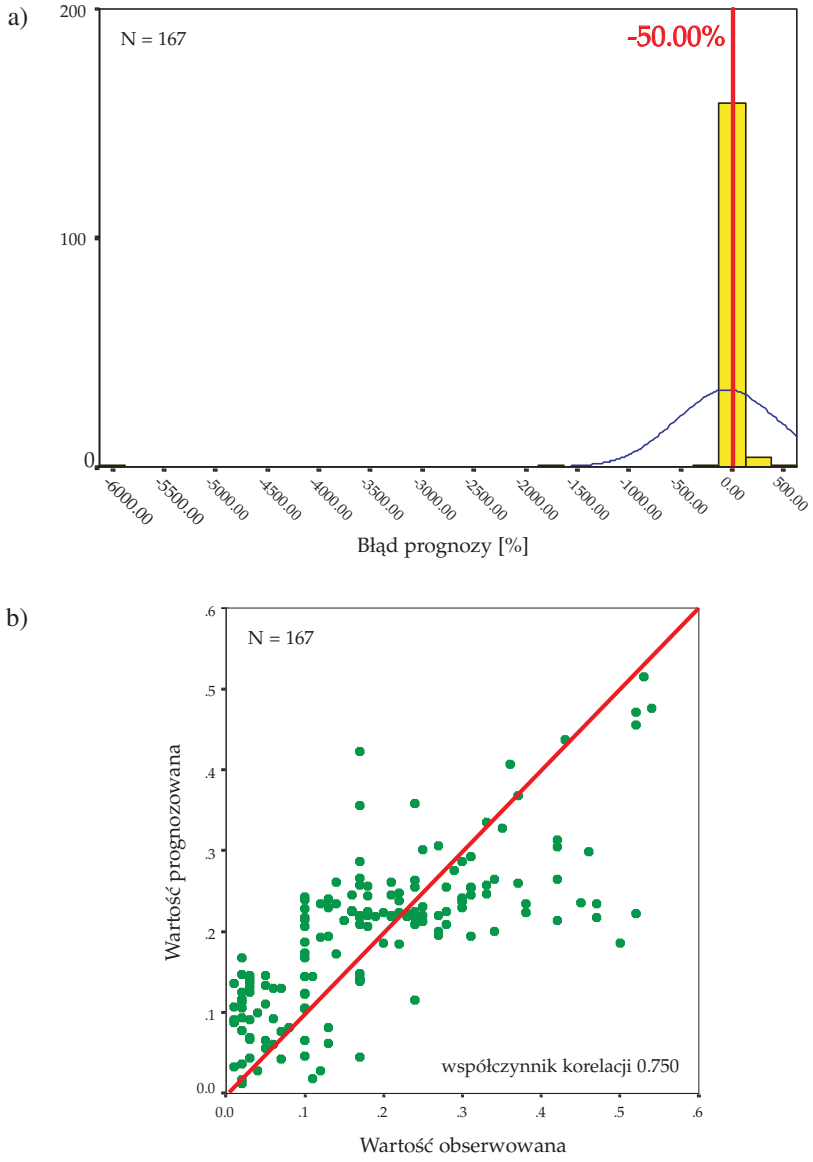


Fig. 32. Prognozowanie stężeń fluorków [mg/dm^3] na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB, C, D

Predictions of fluorides concentrations [mg/dm^3] on the base of monitoring sites coordinates: a) frequency histogram of relative prediction error, b) scatterplot of observed and predicted values; points of AB, C, D class

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

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. 7).

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

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).

Analizie poddano trzy modele sieci: MLP, RBF i Bayesa, podobnie jak w przypadku wariantu 1. Konfiguracje tych modeli (*mlp12.nno–mlp22.nno*, *rbf16.nno–rbf30.nno*) i uzyskane dla zbioru treningowego wyniki prognoz wartości wskaźników fizykochemicznych wód zestawione są w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

Sieć typu MLP. W przypadku sieci MLP, średnie błędy względne prognoz dla 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*; struktura sieci: 2–12–16 odpowiednio: 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).

Sieć typu RBF. Wyniki prognoz wartości wskaźników fizykochemicznych wód uzyskane dla sieci typu RBF kształtują się na poziomie od kilku do 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.

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.

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.

Najlepsze wyniki — najmniejszy średni błąd względny prognoz uzyskano dla sieci typu RBF przy domyślnej konfiguracji modułu (*rbf16.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*) *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*. Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych.

Pełny raport z tej części analizy znajduje się w publikacji elektronicznej (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek), w pliku *output2.spo*. W tabeli 9 zestawione są średnie błędy prognoz *B* (obliczone wg wzoru 4) dla analizowanych zmiennych oraz współczynniki korelacji wartości obserwowanych i prognozowanych.

Tabela 9

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

Analysis of groundwater quality indicators values predictions on the base of monitoring sites coordinates — RBF network, run file, points of AB class (16 target variables)

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* 16 wskaźników fizykochemicznych wód kształtuje się na niskim poziomie, od setnych części procenta do kilkunastu procent. Najmniejszy błąd cechuje prognozy odczynu pH (−0,04%; fig. 33a), a największy — prognozy oznaczeń magnezu (20,56%; fig. 34a).

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).

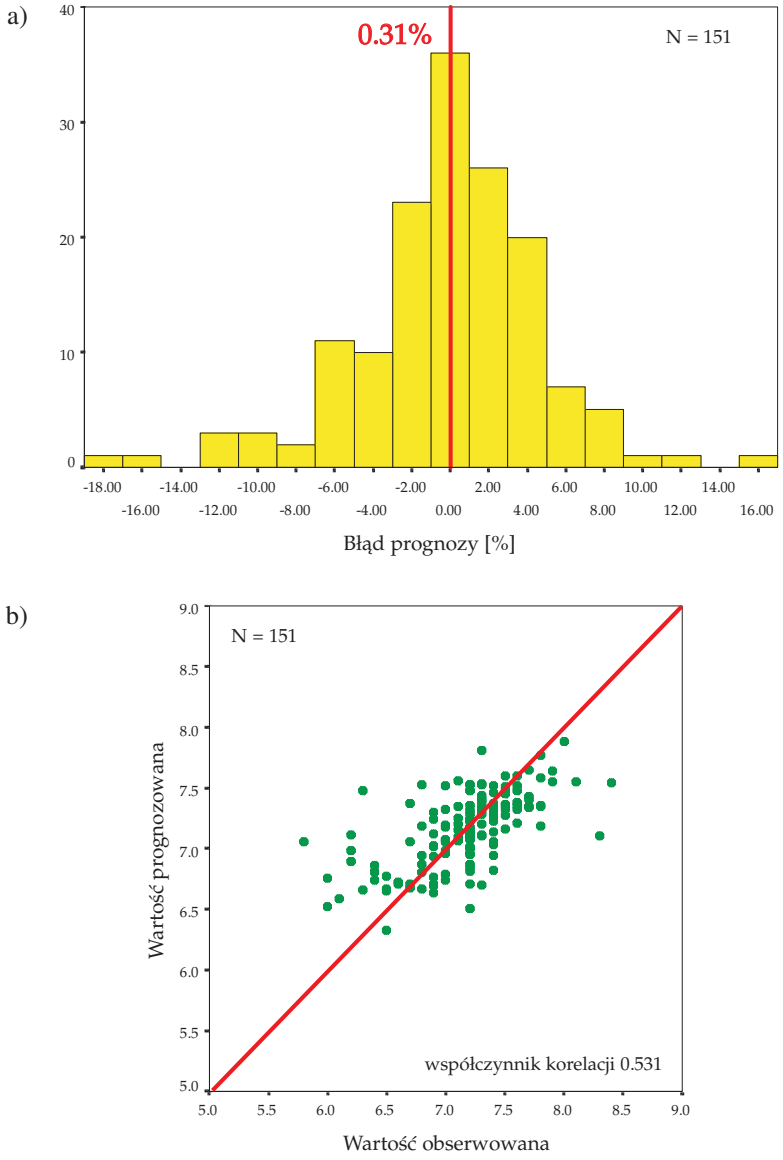


Fig. 33. Prognozowanie pH na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB

Predictions of pH on the base of monitoring sites coordinates: a) frequency histogram of the relative prediction error, b) scatterplot of observed and predicted values; points of AB class

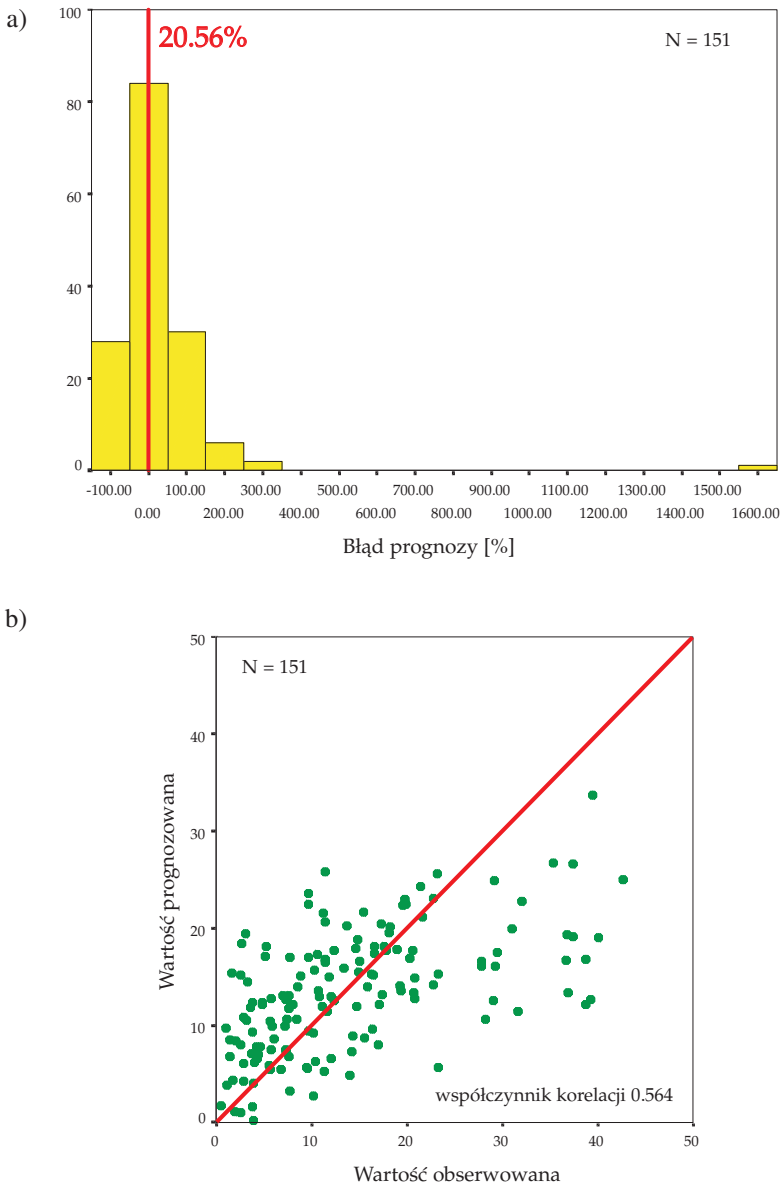


Fig. 34. Prognozowanie stężeń magnezu [mg/dm^3] na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB

Predictions of magnesium concentrations [mg/dm^3] on the base of monitoring sites coordinates: a) frequency histogram of the relative prediction error, b) scatterplot of observed and predicted values; points of AB class

W części przypadków (podobnie jak w wariancie 1) 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.

W przypadku cynku, który charakteryzował się najniższą precyzją ($\sigma^2_{tech} = 86,85\%$) średni błąd względny prognoz B ma wartość 8.67%, obserwuje się jednak duży rozrzut tych błędów w zakresie od -100 do 400%. Współczynnik korelacji wartości obserwowanych i prognozowanych wynosi zaledwie 0,455.

Również w tym przypadku nie stwierdzono związku wielkości błędów uzyskanych prognoz z poziomem wariancji technicznej analizowanych wskaźników (tab. 6). Poziom tych błędów zależy jedynie od konfiguracji sieci.

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. 6), przygotowano kolejny plik testowy (*zbior03.sav*) według wariantu 3 (tab. 7), ograniczony do sześciu zmiennych docelowych (wartości wskaźników fizykochemicznych) bez obserwacji, w których wystąpiły braki danych (143 punkty RMWP).

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. Konfiguracje poszczególnych modeli (*mlp23.nno–mlp33.nno*, *rbf31.nno–rbf45.nno*, *bayes01.nno–bayes08.nno*) i uzyskane dla zbioru treningowego wyniki prognoz wartości wskaźników chemicznych wód zestawione są w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

Sieć typu MLP. Średnie błędy względne prognoz w przypadku zbioru obejmującego punkty RMWP o klasie zagrożenia AB przy ograniczonej liczbie zmiennych (wartości wskaźników fizykochemicznych wód) 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* (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*)).

Sieć typu RBF. 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* (odległość: *City Block*; funkcja nieliniowa: *Inv. Quadratic*; parametr funkcji: 0,1; liczba centrów: 5; rozkład centrów: próbny (*Trial*)).

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).

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* (struktura sieci: 2–4–6; automatycznie generowana liczba neuronów w warstwie ukrytej; grupy parametrów: *Separate Weights and Biases*; wybór modelu: *Most Likely Model*). Sieć ta daje gorsze wyniki prognoz (wyższe błędy względne prognoz MA%) niż sieci MLP i RBF, na poziomie 30–60%.

Wybór najlepszego modelu sieci. Najlepsze wyniki — najmniejszy średni błąd względny prognoz sześciu wskaźników jakości wód uzyskano ponownie dla sieci typu RBF 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*. Następnie sporządzono histogramy rozkładu tych błędów oraz wykresy rozrzutu wartości obserwowanych i prognozowanych. Pełny raport z tej części analizy znajduje się w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/˜ek), w pliku *output3.spo*. W tabeli 10 zestawione są średnie błędy prognoz *B* (obliczone wg wzoru 4) dla analizowanych zmiennych oraz współczynniki korelacji wartości obserwowanych i prognozowanych.

Tabela 10

Analiza prognoz wartości wskaźników jakości wód na podstawie współrzędnych punktu monitoringowego — sieć typu RBF, zbiór roboczy, punkty klasy AB (6 zmiennych docelowych)

Analysis of groundwater quality indicators values predictions on the base of monitoring site coordinates — RBF network, run file, points of AB class (6 target variables)

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%; fig. 35a), a największy — podobnie jak w przypadku zbioru z 16 zmiennymi docelowymi — prognozy oznaczeń magnezu (12,42%; fig. 36a), ma on jednak dwukrotnie mniejszą wartość niż dla zbioru z 16 zmiennymi docelowymi.

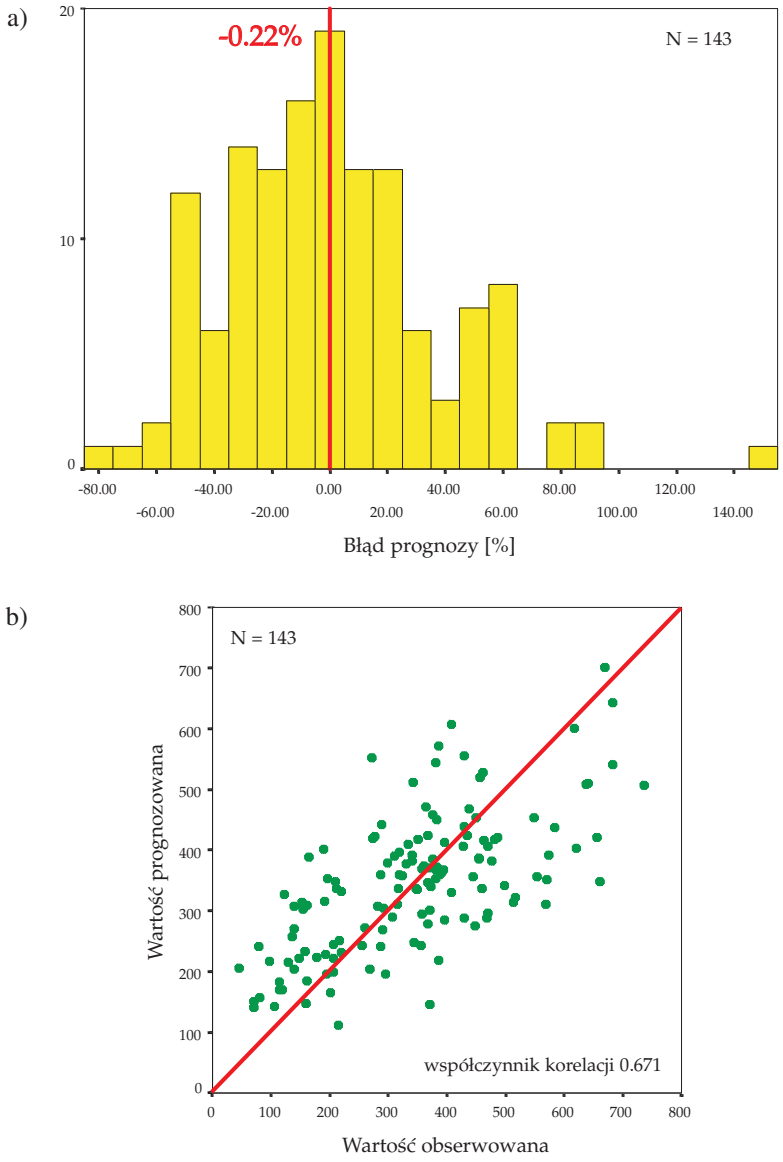


Fig. 35. Prognozowanie stężeń sumy substancji rozpuszczonych [mg/dm^3] na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB

Predictions of total dissolved solids concentrations [mg/dm^3] on the base of monitoring sites coordinates: a) frequency histogram of the relative prediction error, b) scatterplot of observed and predicted values; points of AB class

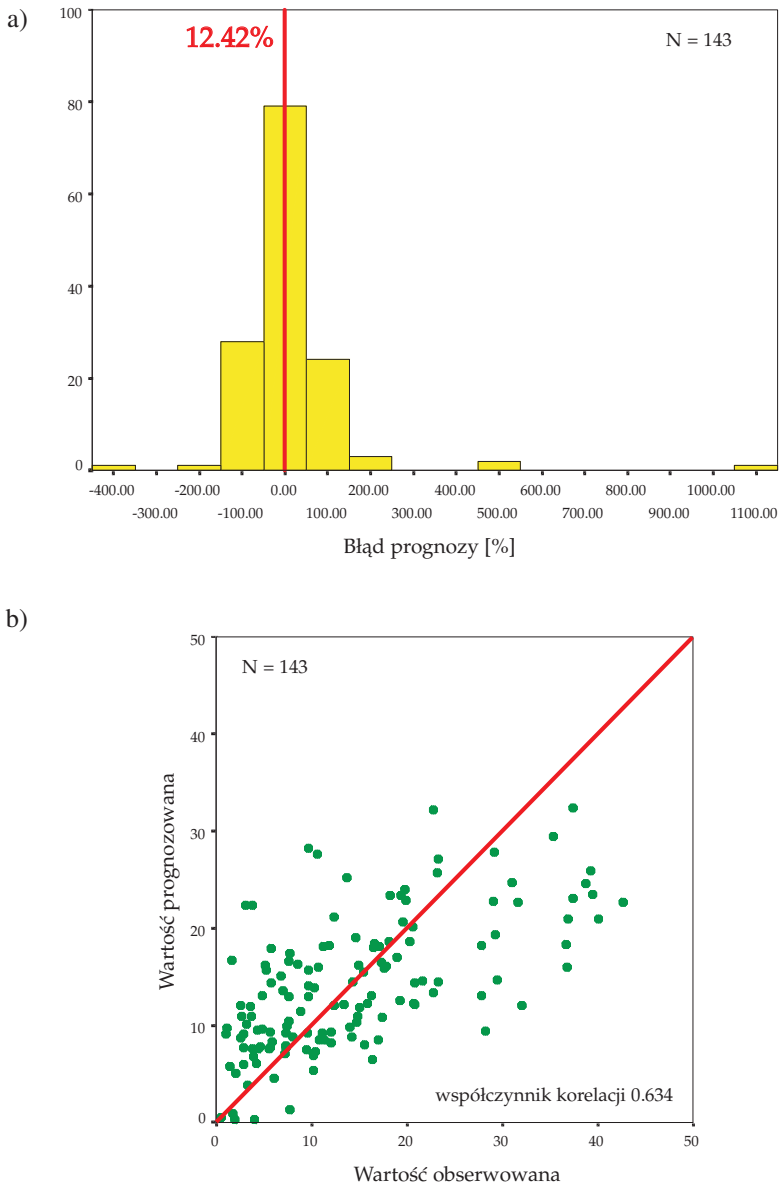


Fig. 36. Prognozowanie stężeń magnezu [mg/dm^3] na podstawie współrzędnych punktu monitoringowego: a) histogram rozkładu błędu względnego prognoz, b) wykres rozrzutu wartości obserwowanych i prognozowanych; punkty klasy AB

Predictions of magnesium concentrations [mg/dm^3] on the base of monitoring sites coordinates: a) frequency histogram of the relative prediction error, b) scatterplot of observed and predicted values; points of AB class

Rozkład błędów względnych prognoz magnezu charakteryzuje się ponadto dużym rozrzutem od –400 do 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. Poziom błędów prognoz zależy jedynie od konfiguracji sieci.

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 wartości oznaczeń wskaźników fizykochemicznych jakości wód podziemnych a zmienną docelową — sposób użytkowania terenu.

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

Przygotowany zgodnie z wariantem pierwszym (tab. 7) plik *zbior01a.sav* (ten plik, tak jak i wszystkie pliki wynikowe oraz pliki z modelami omawianych sieci neuronowych znajdują się w publikacji Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek) wczytano wprost do programu Neural Connection, uruchamiając z programu SPSS opcję **Analiza ▶ Neural Connection**.

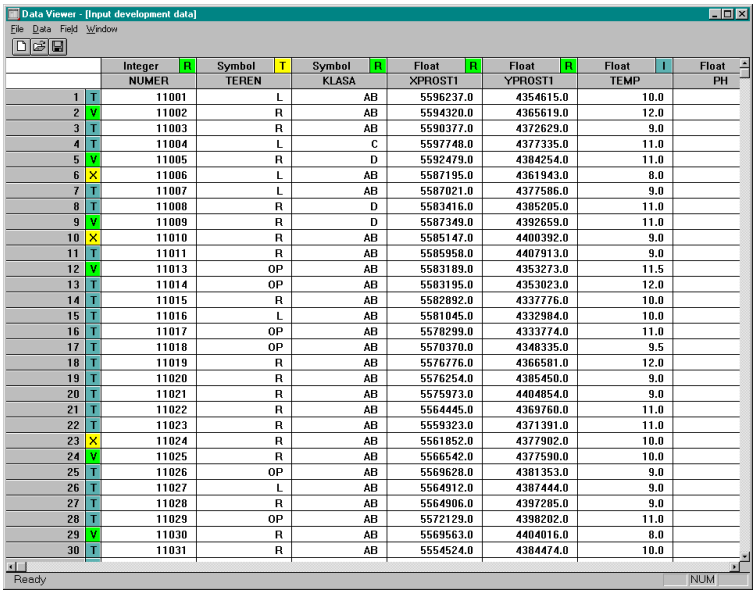
Następnie dokonano konfiguracji zmiennych, w taki sposób, że zmienne: numer identyfikacyjny punktu w bazie MONBADA, klasa zagrożenia wód i współrzędne punktu monitoringowego w układzie 42 zdefiniowano jako zmienne typu opisowego (R), 16 wskaźników fizykochemicznych to zmienne wejściowe (I), a zmienną docelową (T) jest sposób użytkowania terenu w otoczeniu punktu RMWP (fig. 37).

Wskaźniki fizykochemiczne charakteryzujące się rozkładem logarytmiczno-normalnym poddano, za pomocą narzędzia filtrującego, operacji logarytmowania (SPSS, 1997).

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, radialną funkcję bazową RBF, i sieć Bayesa (podobnie jak w przypadku zagadnień predykcji).

Sieć typu MLP. Zbiór danych wejściowych został podzielony na podzbiory (fig. 21): treningowy (80% wszystkich obserwacji — 133 obserwacje), walidacyjny (10% — 17 obserwacji) i testowy (17 obserwacji).



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

Fig. 37. Ekran podglądu danych wejściowych

R — zmienne typu opisowego; I — zmienne typu wejściowego; T — zmienna docelowa

Input data

R — reference variables; I — input variables; T — target variables

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 (pliki *mlp01a–mlp07a*). Uzyskane dla zbioru treningowego wyniki prognoz można znaleźć w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/˜ek).

W tym przypadku parametrem określającym jakość uzyskanych prognoz jest procent obserwacji poprawnie zaklasyfikowanych (punktów monitoringowych poprawnie przyporządkowanych do obszaru o określonym użytkowaniu terenu).

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 klasyfikuje ok. 70% punktów RMWP.

Sieć typu RBF. Podział zbioru danych wejściowych zachowano 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 (pliki *rbf01a–rbf07a*) i uzyskane wyniki prognoz — procent obserwacji poprawnie zaklasyfikowanych — znajdują się w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

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.

Sieć typu Bayesa. Ponieważ sieć Bayesa nie korzysta ze zbioru walidacyjnego, zbiór danych wejściowych podzielono na podzbiory 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 (pliki *bayes01a–bayes07a*) i uzyskane wyniki prognoz — procent obserwacji poprawnie zaklasyfikowanych — zestawione są w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek). Sieć typu Bayesa poprawnie klasyfikuje ok. 90% punktów.

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, najlepszym modelem pozwalającym na klasyfikowanie punktu do obszaru o określonym sposobie zagospodarowania terenu na podstawie wartości wyników oznaczeń wskaźników fizykochemicznych wód (największy procent obserwacji poprawnie zaklasyfikowanych) okazał się model sieci Bayesa, którego wyniki zapisane są w pliku *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*).

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 klasyfikowała 88,6% obserwacji — punktów monitoringowych — ze zbioru roboczego.

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

Kolejne testy przeprowadzono dla zbioru *zbior02a.sav* utworzonego wg wariantu 2 (tab. 7).

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 (pliki *mlp08a–mlp14a*, *rbf08a–rbf14a*, *bayes08a–bayes15a*) i uzyskane dla zbioru treningowego wyniki prognoz — procent obserwacji poprawnie zaklasyfikowanych — zestawione są w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

Porównując wyniki prognoz (procent punktów RMWP klasy AB poprawnie zaklasyfikowanych do obszaru o określonym użytkowaniu terenu) uzyskanych za pomocą modeli MLP, RBF i Bayesa, najlepszym modelem pozwalającym na klasyfikowanie punktu do obszaru o określonym zagospodarowaniu na podstawie wartości wyników oznaczeń wskaźników fizykochemicznych wód (największy procent obserwacji poprawnie zakwalifikowanych) okazał się model sieci Bayesa, którego wyniki zapisane są w pliku *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: *Commitee Decision*).

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 klasyfikowała 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.

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, utworzonego wg wariantu 3 (tab. 7).

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 sześciu 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 zgodnie z wariantami 1 i 2, konfiguracje poszczególnych modeli (pliki *mlp15a–mlp21a*, *rbf15a–rbf20a*, *bayes16a–bayes23a*) i uzyskane dla zbioru treningowego wyniki prognoz — procent obserwacji poprawnie zaklasyfikowanych — zestawione są w publikacji (Kmiecik, 2001; http://galaxy.agh.edu.pl/~ek).

Porównując wyniki prognoz uzyskanych za pomocą modeli MLP, RBF i Bayesa, dla punktów o klasie zagrożenia AB z ograniczoną liczbą zmiennych wejściowych, najlepszym modelem pozwalającym na klasyfikowanie punktu RMWP do obszaru o określonym użytkowaniu terenu na podstawie wartości wyników oznaczeń wskaźników fizykochemicznych wód (największy procent obserwacji poprawnie zakwalifikowanych) okazał się model sieci MLP, którego wyniki zapisane są w pliku *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).

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 klasy-

fikacji 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

Zarówno w Polsce, jak i na świecie zgromadzono znaczne ilości danych uzyskanych w ramach monitoringu jakości wód podziemnych. Klasyczna interpretacja tych danych wskazuje na niski stopień ich wiarygodności, a sama ocena zmienności ma bardzo często charakter formalny i ogranicza się jedynie do ich prezentacji.

Jako obiekt do testowania i wypracowywania zasad kontroli jakości danych w monitoringu oraz oceny zmienności przestrzennej składników fizykochemicznych wód podziemnych wybrano istniejącą bazę danych, zawierającą wyniki uzyskane w ramach regionalnego monitoringu jakości wód podziemnych RMWP przeprowadzonego dla zlewni górnej Wisły w latach 1993–1994 (Witczak i in., 1994a, b).

Lokalizacja punktów RMWP uwzględnia ich reprezentatywność dla określonego typu antropopresji związanej z użytkowaniem terenu.

Informacji o stanie zanieczyszczenia wód pod wpływem antropopresji może dostarczać 90,1% punktów monitoringowych w sieci RMWP dorzecza górnej Wisły. Są to punkty o klasie zagrożenia AB (tzw. wody zagrożone, czas migracji wody z powierzchni terenu do monitorowanej warstwy wodonośnej wynosi do 25 lat). Wody słabo zagrożone (czas migracji 25–100 lat) monitorowane są w 7,0% punktów, a tylko 2,9% stanowią punkty monitoringowe o klasie zagrożenia D (wody praktycznie niezagrożone, czas migracji ponad 100 lat).

Analizie poddano wyniki badań jakości wód podziemnych pobranych w pierwszej serii opróbowania (okres mokry, V–IX 1993). W serii tej opróbowaniem i analizą objęto 167 punktów RMWP — pozostałe punkty (5 punktów RMWP) ze względu na niezakończony proces ich adaptacji nie zostały opróbowane (Witczak i in., 1994a, b).

Przed wprowadzeniem danych do modelu sieci neuronowej wyniki oznaczeń terenowych i laboratoryjnych 55 wskaźników fizykochemicznych (organicznych i nieorganicznych) wód poddano weryfikacji w celu usunięcia danych obarczonych błędami. Błędy w bazie danych wejściowych skutkowałyby powielaniem ich w prognozach dotyczących zmian jakości wód. Podstawę do tej weryfikacji stanowiły wyniki badań przeprowadzonych na próbkach kontrolnych (próbki zerowe i dublowane) pobranych w trakcie terenowego programu kontroli jakości QA/QC prowadzonego równolegle z opróbowaniem sieci monitoringowej RMWP dorzecza górnej Wisły.

Dane hydrogeochemiczne zweryfikowane zostały na trzy sposoby: 1) porównano granice oznaczalności badanych wskaźników: laboratoryjną DL (próbki ślepe) i praktyczną PDL (próbki zerowe); 2) oszacowano wariancję techniczną σ^2_{tech} na podstawie badania próbek dublowanych z wykorzystaniem klasycznej analizy wariancji ANOVA oraz elastycznego postępowania statystycznego (*robust statistics*; program ROB2); 3) dokonano statystycznej analizy rozkładu tych wskaźników z wykorzystaniem programu SPSS (SPPS, 1997a, 2000).

Ze zbioru, na którym oparto prognozowanie zmian jakości wód za pomocą sieci neuronowych usunięto wskaźniki, dla których stosunek PDL/DL miał dużą wartość oraz wskaźniki, dla których wariancja techniczna obliczona metodą klasyczną przekraczała dopuszczalny poziom 20%. Wyłączono też z analizy obserwacje anomalne, obarczone błędami grubymi, i oznaczenia tych parametrów, w zbiorach których ponad 20% stanowiły wyniki poniżej granicy oznaczalności < DL.

W efekcie w zweryfikowanej bazie danych pozostało 16 zmiennych (wskaźników fizykochemicznych): 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].

Na tak przygotowanej bazie danych przeprowadzono próby **predykcji** wartości wskaźników fizykochemicznych wód dla punktu monitoringowego o określonych współrzędnych oraz **klasyfikacji** punktu monitoringowego (na podstawie wartości wyników oznaczeń wskaźników fizykochemicznych) do obszaru o określonym użytkowaniu terenu. Do rozwiązania tych zagadnień wykorzystano modele sieci neuronowych z grupy sieci nadzorowanych (*supervised*): wielowarstwowy perceptron MLP, radialną funkcję bazową RBF i sieć Bayesa. Modele te budowano i testowano w programie Neural Connection (SPSS, 1997).

Badania prowadzono dla trzech wariantów danych zweryfikowanych:

- wariant 1: zbiór zawierający wszystkie zweryfikowane wskaźniki fizykochemiczne (16) i punkty monitoringowe o klasach zagrożenia wód AB, C, D (167 punktów RMWP);
- wariant 2: zbiór zawierający wszystkie zweryfikowane wskaźniki fizykochemiczne (16), ale punkty monitoringowe ograniczone do klasy zagrożenia AB (151 punktów RMWP);
- wariant 3: zbiór zawierający punkty monitoringowe o klasie zagrożenia AB (143 punkty RMWP) i 6 wskaźników zweryfikowanych (spośród zweryfikowanych wskaźników fizykochemicznych do pliku wybrano te, w których wystąpiła najmniejsza liczba braków danych, $n \leqslant 5$).

Różne warianty danych wejściowych umożliwiły ocenę wpływu liczby zweryfikowanych wskaźników fizykochemicznych wód oraz liczby punktów monitoringowych (o różnym stopniu zagrożenia) w bazie danych na jakość uzyskiwanych prognoz. Jakość uzyskiwanych prognoz oceniano na podstawie średniego błędu względnego prognoz MA% obliczanego dla zbiorów treningowych.

Następnie do modelu o najmniejszym względnym błędzie prognoz MA%, w omawianym przypadku we wszystkich wariantach był to model sieci RBF, wprowadzano dane zewnętrzne, tzw. robocze, by sprawdzić jakość uzyskiwanych z modelu prognoz dla 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 jednej klasie zagrożenia AB (warianty 2 i 3) zaobserwowano zmniejszenie wartości średniego błędu względnego prognoz. Współczynniki korelacji wartości obserwowanych z prognozowanymi dla poszczególnych zmiennych (wskaźników fizykochemicznych) kształtowały się na poziomie 0,383–0,904. Poziom błędów uzyskanych prognoz zależał jedynie od konfiguracji sieci.

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 RMWP. 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. Najlepsze wyniki (91,4% obserwacji poprawnie zaklasyfikowanych) uzyskano dla zbioru przygotowanego zgodnie z wariantem 2 (zbiór zawierający wszystkie zweryfikowane wskaźniki fizykochemiczne — 16, ale punkty monitoringowe ograniczone do klasy zagrożenia AB — 151 punktów RMWP).

Uzyskane wyniki badań wskazują zatem, 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.

Podziękowania. Składam serdeczne podziękowania Pani Profesor Jadwidze Szczepańskiej za życzliwość i opiekę naukową podczas realizacji niniejszej pracy. Prof. prof. Aleksandrze Macioszczyk i Zdzisławowi S. Hippe dziękuję za cenne uwagi krytyczne, które uwzględniłam przy przygotowaniu pracy do druku. Dziękuję również Piotrowi Komornickiemu (SPSS Polska) za nieodpłatne udostępnienie programu Neural Connection wraz z dokumentacją, Kaji Gadowskiej za pomoc w przygotowaniu angielskiej wersji streszczenia, a Jacusiowi — za wsparcie TEX-niczne...

LITERATURA

BEDNARCZYK S., 1998 — Metodyka Regionalnego Monitoringu Wód Podziemnych w świetle badań na wybranym obszarze zlewni górnej Wisły. Arch. ZHiOW AGH, Kraków.

BLASCHKE Z., 1995 — Ogólna charakterystyka metody „uogólnionego portretu". W: Materiały XIX sympozjum „Zastosowania metod matematycznych i informatyki w geologii", AGH, Kraków.

BŁASZYK T., MACIOSZCZYK A., 1993 — Klasyfikacja jakości zwykłych wód podziemnych dla potrzeb monitoringu środowiska. Biblioteka Monitoringu Środowiska, Wyd. PIOŚ, Warszawa.

BRODA P., 2000 — Wykorzystanie sieci neuronowych w badaniach geologiczno-geofizycznych. *Wiertnictwo, Nafta, Gaz*, **16**: 143–153.

BRODA P., TWARDOWSKI K., 2001 — Możliwości zastosowania sztucznych sieci neuronowych do określania refleksyjności witrynitu na podstawie wybranych parametrów jakości węgla. XII Międzynarodowa konferencja naukowo-techniczna „Nowe metody i technologie w geologii naftowej, wiertnictwie, eksploatacji otworowej i gazownictwie", 21–22 czerwca 2001, T. 1, AGH, Kraków.

DYNOWSKA J., MACIEJEWSKI M. [red.], 1991 — Dorzecze górnej Wisły. t. I, II. PWN, Warszawa-Kraków.

FALKUS J., PIETRZYKIEWICZ P., 2000 — Zastosowanie sztucznych sieci neuronowych w metalurgii. *Spraw. z Pos. Kom. Nauk. PAN*, **44**, 1: 120.

GLAZOR A., BRODA P., TWARDOWSKI K., 2001 — Porównanie efektywności sztucznych sieci neuronowych z analizą regresji w określaniu przepuszczalności skał na podstawie danych petrofizycznych. XII Międzynarodowa konferencja naukowo-techniczna „Nowe metody i technologie w geologii naftowej, wiertnictwie, eksploatacji otworowej i gazownictwie", 21–22 czerwca 2001. T. 1, AGH, Kraków.

GÓRNIAK J., WACHNICKI J., 2000 — Pierwsze kroki w analizie danych. SPSS Polska, Kraków.

GRUSZCZYŃSKI S., 2000 — Symulacja skutków przekształceń gleb na terenach górniczych za pomocą klasyfikatorów neuronowych. Wyd. AGH, Kraków.

HELSEL D. R., HIRSCH R M., 1992 — Statistical methods in water resources. Elsevier, Amsterdam/London/New York/Tokyo.

HIPPE Z., 2000 — Data Mining and Knowledge Discovery in Chemistry: Possibilities and Limitations. Arch. ZHiOW AGH, Kraków.

HORDEJUK T., 1993 — Krajowy monitoring wód podziemnych — organizacja, główne wyniki prac i badań. W: Biologia i monitoring wód podziemnych (red. E. Kowalczyk, A. Szczepański). 16–17.11.1993, Częstochowa.

HORDEJUK T., GAWIN A., 1994 — Wyniki monitoringu jakości zwykłych wód podziemnych w latach 1991–1993 (sieć krajowa). Biblioteka Monitoringu Środowiska, Wyd. PIOŚ, Warszawa.

JARZYNA J., 2000 — Wyznaczanie przepuszczalności skał na podstawie różnorodnych parametrów petrofizycznych z wykorzystaniem sieci neuronowych. W: Prace Instytutu Górnictwa Naftowego i Gazownictwa nr 110. IGNiG, Kraków.

KALABIŃSKI J., MASTEJ W., 1995 — Komputerowy pakiet obliczeniowy „Metody Rozpoznawania Obrazów". W: Mat. XIX Symp. „Zastosowania metod matematycznych i informatyki w geologii", AGH, Kraków.

KAZIMIERSKI B., SADURSKI A. [red.], 1999 — Monitoring osłonowy ujęć wód podziemnych. Metody badań. Wyd. PIG, Warszawa.

KLECZKOWSKI A. S. [red.], 1990 — Mapa obszarów głównych zbiorników wód podziemnych (GZWP) w Polsce wymagających szczególnej ochrony. Skala 1 : 500 000. Wyd. AGH, Kraków.

KLECZKOWSKI A. S. i in., 1991 — ZTE — Założenia techniczno-ekonomiczne monitoringu jakości wód podziemnych dla dorzecza górnej Wisły. Krakowski region wodnogospodarczy (Kr). Arch. Inst. HiGI, AGH, Kraków (niepubl.).

KMIECIK E., 2000 — Prediction of long-term quality transformations of leachate from coal-mining waste dump with the use of the neural networks. W: International Conference on Exibition and Contaminants in Central and Eastern Europe, Prague 2000 (publikacja na CD-ROM).

KMIECIK E., 2001 — Optymalizacja gęstości opróbowania sieci monitoringowych z wykorzystaniem sieci neuronowych. Arch. ZHiOW AGH, Kraków; publ. elektroniczna na str. http://galaxy.agh.edu.pl/~ek.

KNOSALA R. i in., 2002 — Zastosowania metod sztucznej inteligencji w inżynierii produkcji. WNT, Warszawa.

KOSIŃSKI R. A., 2002 — Sztuczne sieci neuronowe. Dynamika nieliniowa i chaos. WNT, Warszawa.

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: Mat. XIX symp. „Zastosowania metod matematycznych i informatyki w geologii", AGH, Kraków.

KOTLARCZYK J., MASTEJ W., KALABIŃSKI J., 1997 — Wyniki zastosowania nowej strategii rozpoznawania złóż Zn–Pb. W: Mat. XX symp. „Zastosowania metod matematycznych i informatyki w geologii", AGH, Kraków.

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* **15**: 45–68.

KROPKA J., RÓŻKOWSKI A., 1994 — Wstępne wyniki regionalnego monitoringu jakości wód triasowych zbiorników wód podziemnych. W: Zaopatrzenie w wodę miast i wsi (red. M. Sozański). Mat. pokonf. XIII Międzynar. konf. Poznań.

LACHTERMACHER G., FULLER J. D., 1994 — Backpropagation in hydrological times series forecasting. W: Stochastic and statistical methods in hydrology and environmental engineering. vol. 3. „Time series analysis in hydrology and environmental engineering" (eds. K. W. Hipel, A. I. McLeod, U. S. Panu, V. P. Sing). Kluwer Academic Publishers, Dodrecht/Boston/Londyn.

LULA P., 2001 — Wykorzystanie sztucznej inteligencji w prognozowaniu. Statsoft Polska sp. z o.o. (publ. elektroniczna na str. http://www.statsoft.com.pl).

LUSZNIEWICZ A., SŁABY T., 1997 — Statystyka stosowana. PWE, Warszawa.

MACIOSZCZYK A., 1990 — Tło i anomalie hydrogeochemiczne. Metody badania, oceny i interpretacji. CPBP 04.10.09, z. 54. Arch. ZHiOW AGH, Kraków.

MACIOSZCZYK A., DOBRZYŃSKI D., 2003 — Hydrogeochemia. Wyd. PWN, Warszawa.

MASTEJ W., 2001 — Zastosowanie sieci neuronowych do wskazywania zasięgów ciał rudnych w złożu Zn-Pb „Trzebionka" z rejonu śląsko-krakowskiego. W: „Nauki o Ziemi w badaniach podstawowych, złożowych i ochronie środowiska na progu XXI wieku". WGGiOŚ AGH, Kraków.

MCCULLOCH W. S., PITTS W., 1943 — A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**: 115–133.

NABAGŁO I., 1994 — Zastosowanie sieci neuronowych do predykcji nieliniowych sygnałów losowych. W: Materiały konferencyjne I krajowej konferencji „Sieci neuronowe i ich zastosowania". T. II. PC, Częstochowa.

NIELSEN D. M., 1991 — Practical handbook of ground-water monitoring. Lewis Publishers, Chelsea.

OSMĘDA-ERNST E., SZCZEPAŃSKA J., WITCZAK S., 1995 — Praktyczna granica oznaczalności (PQL) jako kryterium jakości opróbowania w monitoringu wód podziemnych. W: „Współczesne problemy hydrogeologii". t. VII. (red. J. Szczepańska, R. Kulma, A. Szczepański). AGH, Kraków.

OSMĘDA-ERNST E., BOBROWSKI A., RZEPECKI T., GAJEWSKA I., KNAP W., 1996 — Znaczenie granic wykrywalności i oznaczalności w analizie mikroskładników wód podziemnych. Mat. konf. VII konferencji „Analityka w służbie geologii i ochrony środowiska". 17–21 czerwca 1996, Szelment.

PARZYCH P., 2001 — Zastosowania sztucznych sieci neuronowych w szacowaniu nieruchomości. W: Nowoczesne technologie w geodezji i inżynierii środowiska, konferencja naukowa z okazji jubileuszu 50-lecia Wydziału Geodezji Górniczej i Inżynierii Środowiska, 21–22 września 2001 r., AGH, Kraków.

PETRIDIS V., KEHAGIAS A., 1998 — Predictive modular neural networks. Applications to Time Series. Kluwer Academic Publishers, Boston/Dodrecht/Londyn.

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. Biblioteka Monitoringu Środowiska, Wyd. PIOŚ, Kielce.

POCIASK-KARTECZKA J. [red.], 1999 — Zastosowania sztucznych sieci neuronowych w hydrologii. Inst. Geografii UJ, Kraków.

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. *Appl. Geochemistry*, Suppl. Issue no 2.

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**.

RÓŻKOWSKI A. i in., 1991 — ZTE – Założenia techniczno-ekonomiczne monitoringu jakości wód podziemnych dla dorzecza górnej Wisły. Katowicki region wodnogospodarczy (Ka). INTERGEO, Sosnowiec (niepubl.).

SIWEK P., 1999 — Chemizm i jakość wód podziemnych serii węglanowej zbiornika triasu gliwickiego w świetle monitoringu regionalnego. Arch. Uniwersytet Śląski, Wydz. Nauk o Ziemi, Katowice.

STANIEWICZ-DUBOIS H., 1991 — Wskazówki metodyczne dotyczące tworzenia regionalnych i lokalnych monitoringów wód podziemnych, wyd. I. Biblioteka Moniotirngu Środowiska. Wyd. PIOŚ, Warszawa.

STANIEWICZ-DUBOIS H., 1995 — Wskazówki metodyczne dotyczące tworzenia regionalnych i lokalnych monitoringów wód podziemnych, wyd. II zmienione. Biblioteka Monitoringu Środowiska. Wyd. PIOŚ, Warszawa.

SZCZEPAŃSKA J., WITCZAK S., POSTAWA A., 1996a — Zastosowanie analizy wariancji (ANOVA) do oceny precyzji wyników badań hydrogeochemicznych. W: Problemy hydrogeologiczne południowo-zachodniej Polski, (red. W. Ciężkowski). UW, Wrocław.

SZCZEPAŃSKA J., WITCZAK S., POSTAWA A., 1996b — Zastosowanie analizy wariancji (ANOVA) do oceny jakości badań w monitoringu wód podziemnych. W: Technika i technologia w ochronie środowiska. I Forum Inżynierii Ekologicznej. (red. I. Wiatr). Ekologia, Lublin-Nałęczów.

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. t. VIII. (red. J. Górski, E. Liszkowska). Poznań.

SZCZEPAŃSKA J., KMIECIK E., 1998 — Statystyczna kontrola jakości danych w monitoringu wód podziemnych. Wyd. AGH, Kraków.

SZCZEPAŃSKA J., KMIECIK E., 2000 — Prognozowanie zmian jakości wód w układzie czasowym z wykorzystaniem sieci neuronowych. W: „Technika i technologia w ochronie środowiska". I Forum Inżynierii Ekologicznej. (red. I. Wiatr). Ekologia, Lublin–Nałęczów.

SZCZEPAŃSKA J., KMIECIK E., 2001 — Wykorzystanie sieci neuronowych do oceny czasu oddziaływania składowiska odpadów górniczych na środowisko wodne. W: I Konferencja „Chemometria — metody i zastosowania", IES, Zakopane.

ŚWIERCZ M., 1994 — Application of neural networks to demand prediction in water distribution networks. Mat. konf. I krajowej konferencji „Sieci neuronowe i ich zastosowania". T. II. Częstochowa.

TADEUSIEWICZ R., 1993 — Sieci neuronowe. Akademicka Oficyna Wydawnicza RM, Warszawa.

TADEUSIEWICZ R., MIKRUT R., 1994 — Sieci neuronowe rozpoznające obrazy. W: „Sieci neuronowe i ich zastosowania" I Kraj. Konferencja 12–15.04.1994.

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, Statsoft.

TADEUSIEWICZ R., 2001 — Wprowadzenie do praktyki stosowania sieci neuronowych. Statsoft Polska sp. z o.o. 2001 (publ. elektroniczna na str. http://www.statsoft.com.pl).

WAKSMUNDZKI T., 1995 — Algorytm LI — „najmniejszego przedziału" — próba zastosowania przy rozpoznawaniu złóż. W: Mat. XIX symp. „Zastosowania metod matematycznych i informatyki w geologii", Kraków.

WIATR I., 1998 — Wstęp do II Forum Inżynierii Ekologicznej „Monitoring Środowiska" (red. I. Wiatr, H. Marczak), Ekologia, Nałęczów.

WITCZAK S. i in., 1993a — 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. Arch. ZHiOW AGH, Kraków.

WITCZAK S. i in., 1993b — The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB). Report for the first quarter of 1993. Arch. ZHiOW AGH, Kraków.

WITCZAK S. i in., 1993c — The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB). Report for the second quarter of 1993. Arch. ZHiOW AGH, Kraków.

WITCZAK S. i in., 1993d — 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. Arch. ZHiOW AGH, Kraków.

WITCZAK S. i in., 1993e — 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, Arch. ZHiOW AGH, Kraków.

WITCZAK S., ADAMCZYK A., 1994 — Katalog wybranych fizycznych i chemicznych wskaźników zanieczyszczeń wód podziemnych i metod ich oznaczania. t. I. Biblioteka Monitoringu Środowiska. Wyd. PIOŚ, Warszawa.

WITCZAK S. i in., 1994a — The Groundwater Quality Monitoring (GQM) of the Upper Vistula River Basin (UVRB). Final Report (Text) + Figures + Tables + Appendices 1, 2. Arch. ZHiOW AGH, Kraków.

WITCZAK S. i in., 1994b — Monitoring jakości wód podziemnych w dorzeczu górnej Wisły — obszar wschodni (badania w zakresie kontroli QA/QC). Arch. ZHiOW AGH, Kraków.

WITCZAK S., ADAMCZYK A., 1995 — Katalog wybranych fizycznych i chemicznych wskaźników wód podziemnych i metod ich oznaczania. t. II. Biblioteka Monitoringu Środowiska. Wyd. PIOŚ, Warszawa.

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. RZGW Katowice, Wydział Nauk o Ziemi UŚ, Katowice.

ZAMORSKA J., 1999 — Prognozowanie wybranych właściwości wody w środowisku naturalnym metodą rozpoznawania obrazów. Arch. ZHiOW AGH, Kraków.

ZHANG S. P., WATANABE H., YAMADA R., 1994 — Prediction of daily water demands by neural networks. W: Stochastic and statistical methods in hydrology and environmental engineering. vol. 3. „Time series analysis in hydrology and environmental engineering" (eds. K. W. Hipel, A. I. McLeod, U. S. Panu, V. P. Sing). Kluwer Academic Publishers, Dodrecht/Boston/Londyn.

ZHU MU-LAN, FUJITA M., HASHIMOTO N., 1994 — Application of neural networks to runoff prediction. W: Stochastic and statistical methods in hydrology and environmental engineering. vol. 3. „Time series analysis in hydrology and environmental engineering" (eds. K. W. Hipel, A. I. McLeod, U. S. Panu, V. P. Sing). Kluwer Academic Publishers, Dodrecht/Boston/Londyn.

DYREKTYWA UNII EUROPEJSKIEJ 98/83/EC — Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption.

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.

DZ. U. NR 203, POZ. 1718 — Rozporządzenie Ministra Zdrowia z dnia 19 listopada 2002 roku, w sprawie wymagań dotyczących jakości wody przeznaczonej do spożycia przez ludzi.

GRID, 1993: Stan środowiska w Polsce. (red. R. Andrzejewski, M. Baranowski). Warszawa, Centrum Informacji o Środowisku.

MONITOR POLSKI Nr 6 z 1991 r., poz. 38. Zarządzenie MOŚZNiL z 1 II 1991 roku w sprawie utworzenia regionalnych zarządów gospodarki wodnej.

PN-EN ISO 17025 — Ogólne wymagania dotyczące kompetencji laboratoriów badawczych i wzorcujących.

PRAWO OCHRONY ŚRODOWISKA WSPÓLNOTY EUROPEJSKIEJ. t. VII. Woda. Warszawa, MOŚZNiL, 1996.

SŁOWNIK HYDROGEOLOGICZNY 2002. Dowgiałło J., Kleczkowski A. S., Macioszczyk T., Różkowski A. [red. nauk.]. PIG, Warszawa.

SPSS, 1997 — Neural Connection 2.0. Applications Guide. User's Guide.

SPSS, 1997a — Dokumentacja do programów SPSS v. 7.5, QI Analyst 3.5 DB.

SPSS, 1999 — Neural Connection 2.1. Update.

SPSS, 2000 — Dokumentacja do programu SPSS v. 10.0 PL.

STATSOFT, 2000: Statystyka w badaniach naukowych. Polska wersja Statistica Neural Networks. Seminaria. Kraków.

SUMMARY

In Poland as well as in other countries a large amount of data obtained from groundwater quality monitoring has been gathered. Interpretation of this data represents low reliability level, and estimation of variability of groundwater quality has often only a formal character limited to data presenting.

As an object of testing and working out the rules of data quality control in monitoring and spatial variability estimation of physicochemical components of ground water the existing database, comprising the results of regional groundwater quality monitoring of the upper Vistula river basin (RGQM) carried out in 1993-1994 (Witczak i in., 1994a, b), was selected.

RGQM network of the upper Vistula river basin consists of 172 monitoring sites, 117 of which lie in the area of then established Regional Council for Water Management (RCWM) Kraków (Monitor Polski, No 6/1991/38), and 55 in RCWM Katowice.

The location of RGQM sites takes into account their representativeness for specific types of anthropopressure connected with the type of land use: 65% of the sites are located in agricultural area (R), 25.5% in forest area (L) and 9.5% in settlement-industrial area (O-P).

90.1% of the monitoring sites of RGQM network of the upper Vistula river basin can provide information about the level of water pollution due to anthropopressure. These are the sites of high endangerment class — AB (i.e. endangered water, water migration time from the ground surface to the monitored water-bearing bed — up to 25 years). The water of low endangerment class — C (migration time from 25 to 100 years) are monitored in 7% of the sites. 2.9% of the sites are of D endangerment class (not endangered water, migration time over 100 years).

The analysis was based on the results of the groundwater quality research, physicochemical indicators of water collected in the first sampling campaign (wet term, from May to September 1993). In this series 167 RGQM sites were sampled and analysed. The remaining 5 sites were not sampled, due to the unfinished process of their adaptation (Witczak *et al.*, 1994a, b).

Before loading data to the neural network model the results of field and laboratory determination of 55 physicochemical (organic and inorganic) water indicators were verified. Data containing errors was

removed to avoid further duplication of mistakes in database input in the prognoses on water quality changes. The basis for the verification were determination results of control samples (i.e. field blank samples and duplicate samples) collected in field quality control program QA/QC conducted parallel to RGQM of the upper Vistula river basin network sampling.

Hydrogeochemical data was verified in three ways: 1) comparison of determination limits: laboratory DL (laboratory blank samples) and practical PDL (field blank samples); 2) estimation of technical variance σ^2_{tech} in the total variance σ^2_{total} on the basis of determination results in duplicate samples employing methods of classical analysis of variance ANOVA and elastic statistical procedure — robust statistics ROB2; 3) statistical analysis of data distribution in SPSS (SPSS, 1997a, 2000).

The indicators for which PLD/DL ratio was high and the indicators for which technical variance calculated by classical method exceeded permissible level of 20% were removed from the set on which the prognosis of water quality changes by neural network were made. Anomalous observations, containing gross errors, and determination of the parameters in the set in which the results under determination limit <DL amounted to over 20% were also excluded from the analysis.

Therefore, in the verified database 16 variables were left: temperature [°C]; power hydrogen pH; total dissolved solids [mg/dm^3]; total alkalinity [mval/dm^3]; total hardness [mg $CaCO_3$/dm^3]; sodium [mg/dm^3]; magnesium [mg/dm^3]; calcium [mg/dm^3]; chlorides [mg/dm^3]; sulfates [mg/dm^3]; silicon dioxide [mg/dm^3]; fluorides [mg/dm^3]; zinc [mg/dm^3]; absorption coefficient UV (A 254); dissolved organic carbon [mg/dm^3]; chemical oxygen demand ChZT-Mn [mg/dm^3].

Predictive trials to provide values of physicochemical water indicators for the monitoring sites with known coordinates and classification of monitoring sites (on the basis of the values of physicochemical indicators) to the area of known type of land use were conducted on a verified database. Three models of neural networks from the group of supervised networks were applied: Multilayered Perceptron MLP, Radial Basis Function RBF and Bayesian Network. These models were created and tested in Neural Connection Program (SPSS, 1997).

Tests were carried out for three variants of verified data:

- Variant 1: set containing all verified physicochemical indicators (16) and monitoring sites characterized by all groundwater endangerment classes AB, C, D (167 RGQM sites);
- Variant 2: set containing all verified physicochemical indicators (16) and monitoring sites restricted to groundwater endangerment class AB (151 RGQM sites);
- Variant 3: set containing 6 verified physicochemical indicators (those with the smallest number of missing values, $n \leqslant 5$) and monitoring sites characterized by groundwater endangerment class AB.

Different variants of input data made possible the estimation of the influence of the verified physicochemical water indicators and the number of monitoring sites (of different endangerment level) in database for the quality of the obtained prognoses. The quality of the prognoses was assessed on the basis of the mean relative prediction error MA% calculated for training set.

Subsequently, in order to check quality of the predictions obtained for new input data, external data ("run data") was loaded to the model of the smallest relative prediction error MA%, which for all variants was RBF neural network model. The mean relative error of the prognoses obtained formed at the level from the hundredth part of percent to a dozen or so percent. This error depended on the selection of input variables. The biggest error was observed in the file with 16 predicted variables and RGQM sites characterised by all classes of groundwater endangerment: AB, C and D (variant 1). After restricting the observation in the input data set to RGQM sites of one endangerment class AB (variants 2 and 3), decrease in value of the mean relative prediction error was observed. Correlation coefficients of the observed and predicted values for individual variables (physicochemical indicators) ranged from 0.383 to 0.904. The level of errors for the predictions obtained depended only on network configuration.

Neural networks were also used for making classification. It was checked whether on the basis of the results of water quality determination, one can obtain data on the type of land use at a given RGQM

site. Similarly to the case of predictions, different models of supervised neural networks (MLP, RBF, Bayesian) were constructed to check in which kind of network the greatest number of monitoring sites would be properly classified to the area of defined type of land use.

Afterwards, in order to check capability of the model, test data ("run data") file was loaded to the best neural network model (Bayesian network, the greatest number of properly classified RGQM sites). In correspondence to selected data variant, neural network properly predicted from 84.9% to 91.4% of observations — monitoring sites — from tested data set. Best results were achieved for the set prepared in accordance to variant 2 (set containing all verified physicochemical indicators — 16 — and monitoring sites restricted to groundwater endangerment class AB — 151 RGQM sites).

The results obtained indicate that the new tool — neural networks — can be successfully applied for spatial prediction of changes in groundwater quality. The condition for reliability of the prognoses is verification of input data loaded to the model.

Translated by Kaja Gadowska