Szymon HOFFMAN
Politechnika Częstochowska, Katedra Chemii, Technologii Wody i Ścieków
ul. Dąbrowskiego 69, 42-200 Częstochowa

Aproksymacja stężeń zanieczyszczeń powietrza za pomocą neuronowych modeli szeregów czasowych

W pracy oceniono możliwości aproksymacji stężeń zanieczyszczeń mierzonych na stacjach monitoringu powietrza. Do predykcji stężeń wykorzystano neuronowe modele szeregów czasowych. Jakość modelowania testowano na rzeczywistych danych pochodzących ze stacji monitoringu powietrza Łódź-Widzew, zarejestrowanych w latach 2004-2008. Analizie poddano względnie kompletny zbiór danych, obejmujący stężenia 6 podstawowych zanieczyszczeń powietrza: O_3, NO_2, NO, PM_{10}, SO_2, CO. Celem badawczym było określenie i porównanie dokładności predykcji stężeń różnych zanieczyszczeń powietrza. Modelowanie przeprowadzono, stosując sztuczne sieci neuronowe. Trening sieci odbywał się przy użyciu liniowego algorytmu pseudoinwersji. Wyjściem modelu było stężenie wybranego zanieczyszczenia w określonym czasie. Wejściami były wartości stężeń zarejestrowane w godzinach wcześniejszych. Każdy model charakteryzowały dwie wielkości: horyzont prognozy i liczba wartości opóźnionych. W analizie określono dokładność predykcji stężeń wybranych zanieczyszczeń dla stałej liczby wartości opóźnionych równej 24 przy zmieniającym się horyzoncie prognozy od 1 do 240 godz. Jako kryterium jakości modelowania przyjęto wartość błędu aproksymacji.

Słowa kluczowe: **szereg czasowy, modele neuronowe, zanieczyszczenia powietrza, monitoring powietrza, stężenia chwilowe, dane monitoringu, brakujące dane, luki pomiarowe, aproksymacja**

Wprowadzenie

Analizę szeregów czasowych można przeprowadzić, stosując klasyczne techniki analizy statystycznej, których celem jest znalezienie jawnej postaci matematycznej dwóch składników szeregów czasowych: trendu i sezonowości [1]. Taka analiza jest żmudna, wymaga spełnienia pewnych warunków, w tym odpowiedniego przygotowania danych, przeprowadzania analiz wstępnych oraz arbitralnego wyboru rodzaju zależności matematycznych. Zastosowanie metody sieci neuronowych do analizy szeregów czasowych pozwala znacznie uprościć procedurę postępowania. Neuronowe modele szeregów czasowych w wielu przypadkach mają zadowalającą dokładność, a ich tworzenie jest szybkie i nie musi być poprzedzane wstępnymi analizami [2-4].

Zbiory wartości stężeń zanieczyszczeń mierzonych w sposób ciągły na stacjach monitoringu powietrza w naturalny sposób tworzą szeregi czasowe. W jednakowych odstępach czasu, co godzinę, rejestrowane są tzw. stężenia chwilowe, będące

w rzeczywistości uśrednionymi stężeniami w okresach 1-godzinnych. Poziomy stężeń zanieczyszczeń powietrza zmieniają się cyklicznie. Charakterystyczne zmiany są obserwowane w cyklu dobowym oraz w cyklu rocznym. Serie czasowe zanieczyszczeń powietrza wykazują autoregresję, która może być wykorzystana w modelowaniu stężeń. Dotychczasowe doświadczenia w zakresie predykcji za pomocą szeregów czasowych pozwalają stwierdzić, że najdokładniejsze modele można uzyskać dla stężeń ozonu - jedynego zanieczyszczenia wtórnego, rejestrowanego na stacjach monitoringu powietrza. W przypadku ozonu prognozowanie stężenia daje zadowalające wyniki nawet dla długiego horyzontu prognozy [5]. W przypadku zanieczyszczeń pierwotnych, takich jak NO_2, NO, PM_{10}, SO_2, CO, dokładność prognozowania szybko maleje z wydłużaniem horyzontu prognozy [6].

W prezentowanej pracy analizie poddano 5-letni, względnie kompletny zbiór danych zarejestrowanych na stacji monitoringu powietrza, obejmujący stężenia 6 podstawowych zanieczyszczeń powietrza: O_3, NO_2, NO, PM_{10}, SO_2, CO. Celem badawczym było określenie i porównanie dokładności predykcji stężeń różnych zanieczyszczeń powietrza przy zastosowaniu metody szeregów czasowych. Aproksymację stężeń przeprowadzono, stosując sztuczne sieci neuronowe.

1. Opis danych i metodyka obliczeń

Przedstawione w pracy badania przeprowadzono, wykorzystując dane pomiarowe zarejestrowane w postaci średnich 1-godzinnych na stacji monitoringu powietrza Widzew w Łodzi w latach 2004-2008. Analizie szeregów czasowych poddano stężenia podstawowych zanieczyszczeń powietrza: O_3, NO_2, NO, PM_{10}, SO_2, CO. Dla serii czasowych każdego z wymienionych zanieczyszczeń wygenerowano oddzielne grupy modeli predykcyjnych.

Do analizy szeregów czasowych wykorzystano liniowe sieci neuronowe. Trening sieci liniowych odbywał się z użyciem algorytmu pseudoinwersji [7]. Wyjściem modelu było stężenie wybranego zanieczyszczenia w określonym czasie. Wejściami były wartości stężeń zarejestrowane w godzinach wcześniejszych. Każdy model charakteryzowały dwie wielkości:

- horyzont prognozy określający liczbę kroków prognozy poza ostatnią z wartości opóźnionych (predyktorów),
- liczba wartości opóźnionych wskazująca, ile przypadków (wcześniejszych obserwacji wartości rozważanej zmiennej) stanowi wejścia sieci.

W analizie określono dokładność predykcji stężeń wybranych zanieczyszczeń dla stałej liczby wartości opóźnionych równej 24, przy zmieniającym się horyzoncie prognozy od 1 do 240 godz. W porównaniach modeli zastosowano następujące kryteria oceny dokładności modeli:

- wartość współczynnika korelacji Pearsona (r),
- wartość pierwiastka z błędu średniokwadratowego (RMSE),
- wartość średniego błędu bezwzględnego (|e|),
- stosunek pierwiastka z błędu średniokwadratowego do odchylenia standardowego (RMSE/s).

Wartości RMSE i |e| mogą być użytecznym kryterium przy porównywaniu sieci modelujących stężenie tego samego zanieczyszczenia na wybranej stacji monitoringu. Kryteriami bardziej uniwersalnymi są współczynnik korelacji i stosunek RMSE/s. Dwie ostatnie miary błędu modelowania umożliwiają porównanie dokładności predykcji sieci modelujących stężenia różnych zanieczyszczeń.

Obliczenia przeprowadzono, korzystając z programu STATISTICA Neural Networks. W przypadku każdej sieci neuronowej zbiór wszystkich przypadków został losowo podzielony na trzy podzbiory: zbiór uczący (50% przypadków), zbiór weryfikujący (25% przypadków), zbiór testujący (25% przypadków).

2. Wyniki badań

2.1. Wstępna ocena danych

Podstawowe parametry statystyczne analizowanych serii czasowych przedstawiono w tabeli 1. Kompletność danych dla prawie wszystkich zanieczyszczeń przekracza 90%, tylko dla NO wynosi 72,1%. Zamieszczone parametry statystyczne zawierają wartość średnią, która w rozkładach wartości stężeń poszczególnych gazów może być traktowana jako miara centralna. Odchylenia standardowe, wartości minimalne i maksymalne charakteryzują zmienność poszczególnych serii czasowych.

Tabela 1

Podstawowe parametry statystyczne analizowanych serii czasowych 1-godzinnych stężeń zanieczyszczeń, Łódź-Widzew 2004-2008

Parametr	Jednostka	O_3	NO	NO_2	CO	SO_2	PM_{10}
Kompletność serii	%	97,1	72,1	91,4	97,5	98,1	93,7
Wartość minimalna	$\mu g/m^3$	0,9	0,0	0,0	20,7	0,0	0,0
Wartość maksymalna	$\mu g/m^3$	197,9	323,3	170,1	5794,8	471,2	264,0
Średnia	$\mu g/m^3$	58,1	3,3	18,1	446,8	14,8	22,7
Odchylenie standardowe	$\mu g/m^3$	31,2	8,4	13,2	240,9	14,2	16,3

2.2. Wyniki modelowania

W tabelach 2-7 zamieszczono wyniki pozwalające określić jakość modelowania mierzonych zanieczyszczeń powietrza. We wszystkich opisanych w tych tabelach modelach szeregów czasowych przyjęto liczbę wartości opóźnionych równą 24, co oznacza, że w predykcji wykorzystano stężenia z pełnej doby pomiarowej. Tabele zawierają wartości czterech różnych miar błędu modelowania i pozwalają obserwować zmiany dokładności predykcji przy wydłużaniu horyzontu prognozy od 1 godziny do 240 godzin.

Tabela 2

Ocena dokładności modelowania chwilowego stężenia O_3 w zależności od przyjętej liczby kroków do prognozowanej obserwacji; stężenia 1-godzinne, Łódź-Widzew 2004-2008; liczba wartości opóźnionych = 24

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	7,96	5,26	0,255	0,967
2	12,15	8,59	0,390	0,921
3	14,79	10,82	0,474	0,880
4	16,52	12,32	0,530	0,848
5	17,69	13,36	0,567	0,823
6	18,50	14,10	0,593	0,805
8	19,45	14,96	0,624	0,782
12	20,24	15,71	0,649	0,761
24	20,29	15,84	0,651	0,759
72	22,26	17,59	0,714	0,700
240	24,68	19,58	0,791	0,612

Tabela 3

Ocena dokładności modelowania chwilowego stężenia NO w zależności od przyjętej liczby kroków do prognozowanej obserwacji; stężenia 1-godzinne, Łódź-Widzew 2004-2008; liczba wartości opóźnionych = 24

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	5,59	1,56	0,668	0,745
2	7,45	2,20	0,890	0,458
3	8,01	2,54	0,957	0,290
4	8,17	2,69	0,977	0,214
5	8,24	2,76	0,984	0,176
6	8,26	2,79	0,987	0,160
8	8,28	2,83	0,990	0,144
12	8,30	2,86	0,993	0,123
24	8,32	2,87	0,994	0,110
72	8,34	2,88	0,997	0,073
240	8,33	2,89	0,996	0,090

Tabela 4.

Ocena dokładności modelowania chwilowego stężenia NO_2 w zależności od przyjętej liczby kroków do prognozowanej obserwacji; stężenia 1-godzinne, Łódź-Widzew 2004-2008; liczba wartości opóźnionych = 24

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	5,84	3,73	0,444	0,896
2	8,39	5,50	0,637	0,770
3	9,79	6,56	0,744	0,668
4	10,55	7,17	0,801	0,598
5	10,97	7,51	0,833	0,553
6	11,21	7,72	0,852	0,524
8	11,46	7,94	0,870	0,492
12	11,72	8,18	0,890	0,455
24	11,93	8,38	0,906	0,423
72	12,47	8,87	0,948	0,319
240	12,60	9,01	0,957	0,290

Tabela 5

Ocena dokładności modelowania chwilowego stężenia CO w zależności od przyjętej liczby kroków do prognozowanej obserwacji; stężenia 1-godzinne, Łódź-Widzew 2004-2008; liczba wartości opóźnionych = 24

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	92,3	45,5	0,383	0,924
2	135,8	70,7	0,564	0,826
3	156,6	85,2	0,650	0,760
4	167,4	93,5	0,695	0,719
5	173,7	99,3	0,721	0,693
6	177,6	102,6	0,737	0,676
8	182,3	107,5	0,757	0,654
12	188,5	113,5	0,783	0,623
24	198,1	121,6	0,822	0,569
72	210,8	133,5	0,875	0,484
240	222,5	146,1	0,927	0,376

Tabela 6

Ocena dokładności modelowania chwilowego stężenia SO_2 w zależności od przyjętej liczby kroków do prognozowanej obserwacji; stężenia 1-godzinne, Łódź-Widzew 2004-2008; liczba wartości opóźnionych = 24

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	7,70	3,70	0,541	0,841
2	9,73	5,03	0,683	0,730
3	10,51	5,62	0,738	0,675
4	10,89	5,93	0,765	0,644
5	11,11	6,14	0,781	0,625
6	11,27	6,29	0,792	0,611
8	11,44	6,45	0,804	0,595
12	11,70	6,66	0,822	0,570
24	12,08	6,98	0,849	0,529
72	12,68	7,38	0,891	0,455
240	13,14	7,77	0,922	0,386

Tabela 7

Ocena dokładności modelowania chwilowego stężenia PM_{10} w zależności od przyjętej liczby kroków do prognozowanej obserwacji; stężenia 1-godzinne, Łódź-Widzew 2004-2008; liczba wartości opóźnionych = 24

Horyzont prognozy godz.	Błąd modelowania			
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE/s	r
1	7,70	4,72	0,474	0,881
2	9,95	6,40	0,612	0,791
3	11,09	7,33	0,682	0,731
4	11,78	7,92	0,725	0,689
5	12,26	8,34	0,754	0,657
6	12,63	8,64	0,777	0,629
8	13,10	9,01	0,806	0,592
12	13,70	9,50	0,843	0,538
24	14,66	10,22	0,902	0,433
72	15,63	11,02	0,961	0,276
240	16,14	11,39	0,994	0,113

3. Dyskusja wyników i podsumowanie

Modele stężeń zanieczyszczeń powietrza na ogół cechuje nie najlepsza dokładność. Brak precyzji wynika z samej koncepcji modelu. W modelowaniu opartym na analizie szeregów czasowych wykorzystuje się wiedzę ukrytą w danych historycznych. Model może pokazać tylko statystyczną zmienność stężeń w krótszym lub dłuższym okresie czasu. Nie uwzględnia jednak ważnych przyczyn, mogących modyfikować stężenia zanieczyszczeń. Najlepsze rezultaty modelowania można uzyskać dla stężeń O_3, zanieczyszczenia wtórnego, charakteryzującego się dość regularną zmiennością dobową stężeń. Stężenia zanieczyszczeń pierwotnych, takich jak NO_2, NO, PM_{10}, SO_2, CO, wykazują mniejszą regularność, ponieważ ich stężenia silnie zależą od emisji lokalnej. Dlatego modelowanie stężeń tych zanieczyszczeń jest obarczone wyższym błędem.

Wyniki przeprowadzonej analizy pokazują, że jakość modelowania silnie zależy od horyzontu prognozy. W przypadku błędów RMSE, |e|, RMSE/s wartości stopniowo rosną w miarę wydłużania horyzontu prognozy. Odmiennie zachowują się wartości współczynnika korelacji, które stopniowo maleją. Z punktu widzenia oceny jakości modelowania wszystkie miary błędu mają podobną charakterystykę - wydłużanie horyzontu prognozy powoduje wzrost błędu.

Miarą błędu, która umożliwia porównanie jakości modeli utworzonych dla różnych zanieczyszczeń, jest współczynnik korelacji r. Korzystając z wcześniejszych doświadczeń, w pracy przyjęto następującą klasyfikację modeli ze względu na ich zdolności predykcyjne [8]:

- modele bardzo dokładne, gdy $r > 0{,}90$
- modele dokładne, gdy $0{,}90 > r > 0{,}80$
- modele mało dokładne, gdy $0{,}80 > r > 0{,}50$
- modele niedokładne, gdy $r < 0{,}50$

Tę klasyfikację można stosować do oceny dokładności modeli autonomicznych, tj. modeli wykorzystujących wyłącznie dane rejestrowane na stacjach monitoringu powietrza [6]. Stosując przedstawioną wyżej klasyfikację, można potwierdzić znany z wcześniejszych publikacji fakt, że najdokładniejsze modele szeregów czasowych otrzymuje się w przypadku ozonu (tab. 2). Modele bardzo dokładne uzyskano przy horyzoncie prognozy wynoszącym do 2 godzin. CO jest jedynym zanieczyszczeniem poza ozonem, dla którego dało się wygenerować model bardzo dokładny, ale wyłącznie dla najkrótszego horyzontu prognozy, równego 1 godz. (tab. 5). Dla pozostałych zanieczyszczeń nie udało się utworzyć modeli bardzo dokładnych. Modele dokładne wygenerowano dla O_3 (horyzont prognozy w zakresie 3-6 godz.), dla CO (horyzont prognozy równy 2 godz.) oraz dla NO_2, SO_2 i PM_{10} (horyzont prognozy równy 1 godz.). Tylko dla NO nie udało się utworzyć modelu dokładnego (tab. 3). Ta obserwacja jest zgodna z wcześniejszymi doniesieniami o trudnościach w modelowaniu stężenia NO za pomocą analizy szeregów czasowych [6, 9].

Analiza szeregów czasowych za pomocą sieci neuronowych może być użytecznym narzędziem modelowania stężeń zanieczyszczeń powietrza przy powierzchni ziemi i może być wykorzystana do uzupełniania brakujących danych w systemach

monitoringu powietrza. Podsumowując wyniki, można stwierdzić, że jakość modelowania podstawowych zanieczyszczeń powietrza maleje w szeregu: O_3, CO, SO_2, PM_{10}, NO_2, NO. Rozważając możliwości wykorzystania analizy szeregów czasowych, jako metody pozwalającej modelować stężenia zanieczyszczeń mierzonych na stacjach monitoringu powietrza, należy podkreślić, że metoda ta nie zapewnia jednakowej jakości modelowania dla różnych zanieczyszczeń. Stosunkowo dokładne modele można uzyskać dla ozonu, w przypadku innych zanieczyszczeń błędy modelowania są większe i bardzo szybko rosną w miarę wydłużania horyzontu prognozy.

Wnioski

Na podstawie przeprowadzonych badań można sformułować następujące wnioski ogólne:

1. Jakość modelowania stężeń zanieczyszczeń powietrza za pomocą sieci neuronowych wykorzystujących analizę szeregów czasowych silnie zależy od horyzontu prognozy. Dokładność aproksymacji stężeń zanieczyszczeń powietrza metodą analizy szeregów czasowych maleje przy wydłużaniu horyzontu prognozy.
2. Dokładność predykcji stężeń podstawowych zanieczyszczeń powietrza maleje w szeregu: O_3, CO, SO_2, PM_{10}, NO_2, NO.
3. Aproksymacja stężenia ozonu daje zadowalające wyniki nawet dla długiego horyzontu prognozy.

Praca naukowa finansowana ze środków na naukę w latach 2006-2008 jako projekt badawczy nr 1 T09D 037 30.

Literatura

[1] Brockwell P.J., Davis R.A., Introduction to time series and forecasting, Springer-Verlag 2002.

[2] Gardner M.W., Dorling S.R., Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences, Atmos. Environ. 1998, 32, 14/15, 2627-2636.

[3] Ballester E.B. i in., Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks, Ecological Modelling 2002, 156, 27-41.

[4] Hoffman S., Jasiński R., Studies on NO_x concentration modeling in the air monitoring systems, chapter in Pathways of pollutants and mitigation strategies of their impact on the ecosystems, ed. M.R. Dudzińska, M. Pawłowska, Monografie Komitetu Inżynierii Środowiska PAN, Lublin 2004, 185-192.

[5] Hoffman S., Missing data completing in the air monitoring systems by forward and backward prognosis methods, Environmental Protection Engineering 2006, 32(4), 25-29.

[6] Hoffman S., Treating missing data at air monitoring stations [w:] Environmental Engineering, Eds. L. Pawłowski, M.R. Dudzińska, A. Pawłowski, Taylor & Francis Group, London 2007, 349--353.

[7] Hoffman S., Zastosowanie sieci neuronowych w modelowaniu regresyjnym stężeń zanieczyszczeń powietrza, Wydawnictwa Politechniki Częstochowskiej, Częstochowa 2004.

[8] Statistica Neural Networks, StatSoft 1998.

[9] Hoffman S., Short-time forecasting of atmospheric NO_x concentration by neural networks, Environmental Engineering Science 2006, 23(4), 603-609.

Aproximation of Air Monitoring Data Gaps by Means of Time-Series Neural Models

An assessment of quality of air pollutants concentration modeling was the main research purpose. The examination was made by means of artificial neural networks, which were employed to create time-series models. The quality of approximation was tested on the actual set of air monitoring data, gathered over a 5-year period at the measure site in Lodz-Widzew (Central Poland). The examined time-series involved hourly concentrations of main air pollutants: O_3, NO_2, NO, PM_{10}, SO_2, CO. The research aim was the estimation and the comparison of prediction accuracy for different air pollutants. Time-series models were characterized by two parameters which might influence the prediction quality: lookahead and steps. For all models the constant number of steps equal 24 hours was assumed. The effect of changes of lookahead in the range 1÷240 hours was analyzed. It was stated that the decreasing of precision of time-series models with the increase of lookahead is observed. The drop of accuracy depends on pollutant. The furthest reasonable prognosis may be done for ozone concentration. Approximation accuracy shortens in the order: O_3, CO, SO_2, PM_{10}, NO_2, NO.

Keywords: time series, neural models, air pollution, air monitoring, hourly concentrations, monitoring data, missing data, measure gaps, approximation